Full Title: Evolution is in the details: Regulatory differences in modern human and Neanderthal
Short title: Regulatory differences in modern human and Neanderthal

Harlan R. Barker [1,2*], Seppo Parkkila [1,2], Martti E.E. Tolvanen [3]

1 Faculty of Medicine and Health Technology, Tampere University, 33520 Tampere, Finland.

2 Fimlab Ltd., Tampere University Hospital, 33520 Tampere, Finland.

3 Department of Computing, University of Turku, 20014 Turku, Finland.

*To whom correspondence should be addressed.

Abstract

Transcription factor proteins play a critical role in the regulation of eukaryote gene expression via sequence-specific binding to genomic locations known as transcription factor binding sites.

We studied sites of genomic variation between modern human and Neanderthal promoters. We detected significant differences in the binding affinities of 110 transcription factors to the promoters of 74 target genes. The transcription factors were enriched for terms related to vision, motor neurons, homeobox, and brain, whereas the target genes and their direct interactors were enriched in terms related to autism, brain, connective tissue, trachea, prostate, skull morphology, and vision. Secondary analysis of single-cell data revealed that a subset of the identified transcription factors (CUX1, CUX2, ESRRG, FOXP1, FOXP2, MEF2C, POU6F2, PRRX1 and RORA) co-occur as marker genes in L4 glutamatergic neurons. The majority of these genes have known roles in autism and/or schizophrenia and are associated with human accelerated regions (elevated divergence in humans vs. other primates). These results support the value of gene regulation studies for the evolution of human cognitive abilities and the neuropsychiatric disorders that accompany it.

Introduction

North of Eden

Humans and their hominin relatives have been leaving Africa in waves for the past two million plus years. Various environmental and cultural pressures have impacted each diaspora in different ways, subsequently producing adaptations reflected in physiology, immunity, and brain size. Recently, the discovery and sequencing of DNA from the remains of Neanderthal [1–4] and Denisovan [5,6] have allowed direct comparisons of DNA from modern and ancient hominids. Among the observed genomic variations between modern humans and Neanderthals, a limited number have been identified in gene coding regions. Some of these genes are known to affect cognition and morphology (cranium, rib, dentition, and shoulder joint) [1], pigmentation and behavioral traits [2], and brain development [7]. However, as has been noted before, there is a paucity of coding variations to explain the differences between related species; the genome of the Altai Neanderthal reveals just 96 fixed amino acid substitutions, occurring in 87 proteins [7]. Unsurprisingly, a much larger set of variants are observed in intergenic regions, owing not only to the fact that these are comparatively much larger regions but also to the expectation of less conservation in what until recently was often termed "junk DNA". While variants in coding regions can directly affect protein structure, those found in intergenic regions may affect the regulation of gene expression through alternative binding of transcription factors in promoters and enhancers and the expression of noncoding RNAs. What may surprise some is the cumulative effect of numerous small — and large — changes to gene expression arising owing to these manifold intergenic changes, which may ultimately serve as the engine of speciation. Indeed, introgressed Neanderthal DNA is more depleted in regulatory regions than in those encoding proteins [8]. Using a new computational tool that incorporates numerous transcription-relevant genomic features, we sought to reveal how the comparative differences in these regions may affect regulatory differences between modern humans and Neanderthals. Specifically, our aim was to identify those gene-regulating transcription factors whose binding to DNA may vary between these species of hominids and thus drive the differences between them.

Methods

Analysis of modern human vs. Neanderthal genetic variation

The Neanderthal Genome Project has cataloged a total of 388,388 SNPs in modern human and Neanderthal genomes [1]. These were further reduced to 21,990 SNPs in the proximal promoters (-2,500 to +2,500 nts relative to the TSS) of those transcripts, which are defined by Ensembl as "protein-coding" (Figure 1). Using TFBSFootprinter [9–12], a tool we have developed for the prediction of TFBSs, a 50 bp region centered on each SNP was analyzed for the binding of 575 TFs for both the modern human version and the Neanderthal variant. TFBSFootprinter automatically retrieves the human sequences at a target region, and custom Python scripts were used to modify these sequences for the Neanderthal variant. All TFBS predictions that overlapped with the target SNP position were retained. The complete result set was then reduced on the basis of the combined affinity score p value via a Benjamini–Hochberg-derived critical p value corresponding to a false discovery rate (FDR) cutoff of 0.01 to address multiple testing. For each putative TFBS meeting the cutoff in either subspecies, the corresponding matched pair of PWM scores was retained. To identify significantly different scoring for TFs between subspecies, for each TF, via the compiled matched scores, the Wilcoxon rank test was performed via the SciPy stats Python library [13], and subsequent results were filtered via a Benjamini–Hochberg-derived critical p value corresponding to an FDR cutoff of 0.10 to address multiple testing.

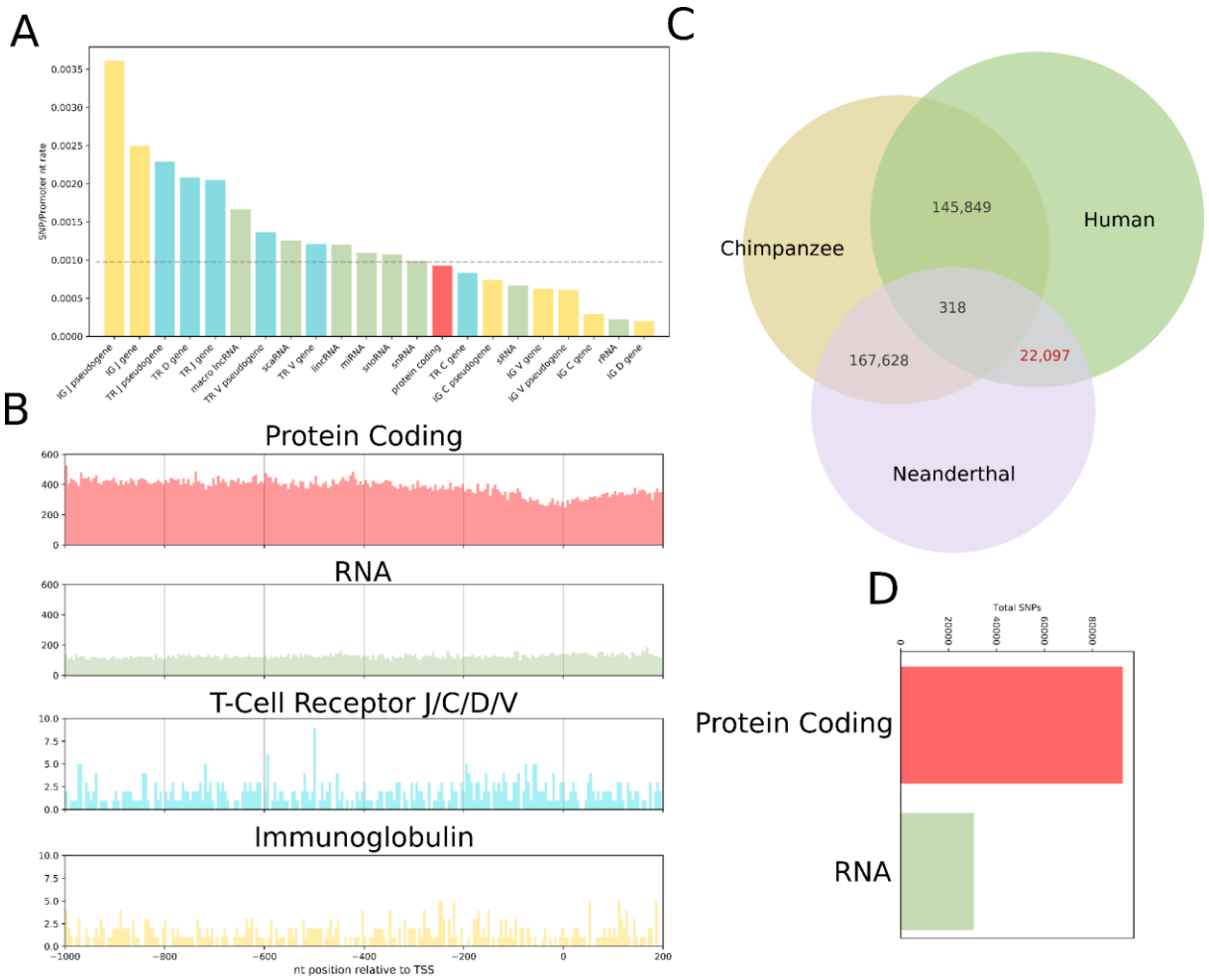


Figure 1. Modern human and Neanderthal variants used in the analysis. (**A**) Incidence rate of SNPs/nt in human promoters of the various transcript classes, as defined by Ensembl. The bars are colored by transcript class and correspond with those in panel B. (**B**) Number of observed modern human/Neanderthal SNPs at each nt location in the sets of promoters of various transcript types. A visible depletion of SNPs is observed for the promoters of protein-coding transcripts corresponding with increasing proximity to the TSS (nt position 0). (**C**) Counts of total SNPs in promoters of protein-coding transcripts between the three species cataloged in the Neanderthal Genome Project. (**D**) Comparison of counts of human vs. chimp SNPs in promoters of protein-coding vs RNA transcripts. IG J, immunoglobulin joining chain; TR J, T-cell receptor joining chain; TR D, T-cell receptor diversity chain; TR V, T-cell receptor variable chain; IG C, immunoglobulin constant chain; IG V, immunoglobulin variable chain; IG D, immunoglobulin diversity chain.

A total of 108 TF models were identified as scoring differently across human vs. Neanderthal SNP locations. For each of these genes, we extracted RNA-Seq data from the FANTOM dataset (across all CAGE peaks associated with that TF) and retained the data for the 100 tissues with the highest aggregate expression across all of the target TF genes. In the case of heterodimer JASPAR TF models (e.g., FOS::JUN, NR1H3::RXRA, and POU5F1::SOX2), the expression of each TF component gene was used. The expression data were extracted as TPM values and normalized by $\log_2$ transformation. From the subsequent normalized expression data values, hierarchical clustering was performed and visualized via SciPy, Matplotlib, and Seaborn Python libraries [13–15].

The results of hierarchical clustering revealed a cluster of TF genes that were uniquely expressed in neural and immune tissues. These gene sets were used to perform PANTHER-based gene ontology enrichment analysis (using the www.geneontology.org web server) [16,17], with the default statistical settings using Fisher's exact test method and an FDR threshold of $p<0.05$ (default setting).

Enrichment analysis of DB TFs and their target genes

The OmniPath DB compiles protein annotation data from several tens of publicly available datasets. The complete set of annotation data was downloaded and filtered to annotations from PanglaoDB and UniProt tissue, family, and keyword classifications. The set of 110 DB TFs was analyzed for enrichment for all annotation labels via Fisher’s exact test and compared with the full set of 529 human JASPAR TF genes, with Benjamini–Hochberg correction for multiple testing (FDR cutoff value of 0.1).

The OmniPath DB has likewise compiled protein interactions (human, mouse and rat) from many publicly available datasets. The complete set of interactions was downloaded and filtered to determine posttranslational human protein–protein interactions. A total of 805 interactors were identified for the set of 74 DB TF target genes, and the combined list (879 genes) was analyzed for gene ontology enrichment analysis via the geneontology.org webserver [16,17], with the default settings of Fisher’s exact test and an FDR cutoff of 0.05 for multiple testing correction.

Analysis of the brain expression of differentially binding TFs

Expression data in the form of reads per kilobase per million reads (RPKM) were extracted for 26 unique tissues across 31 timepoints (from 8 weeks postconception to 40 years after birth) from the Allen Brain Atlas (brainspan.org). The RPKM values were then converted to TPM values and $\log_2$ transformed. Ages were grouped into three phases of growth for simplicity of analysis and interpretation: pcw (8–37 weeks post conception), early (4 months–4 years), and late (8 years–40 years). Corresponding to these age groups, log-transformed TPM values for each tissue were grouped and averaged and used to perform cluster map analysis to identify grouping tissues at time points with similar expression profiles.

Analysis of brain samples via scRNA-Seq

The Allen Brain Atlas has performed single-nucleus RNA-Seq analysis of 49,417 nuclei derived from 8 brain cortex regions within the middle temporal gyrus (MTG), anterior cingulate gyrus (CgG), primary visual cortex (V1C), primary motor cortex (M1C), primary somatosensory cortex (S1C) and primary auditory cortex (A1C) [18]. These data were downloaded as a count matrix along with a table of the associated metadata (https://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq) and loaded into the Python library SCANPY [19]. Using a modified workflow described previously [20], samples were filtered by Gaussian fit of the read count (300,000<x<3,500,000), expressed gene count (2,000<x), and number of cells in which a gene was expressed (>50), resulting in a final count of 46,959 cells and 42,185 genes for further analysis.

Using SCANPY, counts were normalized by cell (‘pp.normalize_total’; target_sum=1,000,000), log transformed (‘pp.log1p’), highly variable genes identified (‘pp.highly_variable_genes’; flavor=‘seurat_v3’; n_top_genes=5,000; layer='counts'), principal component analysis (‘pp.pca’; n_comps = 15; svd_solver='arpack'), and k-nearest neighbors (‘pp.neighbors’; n_neighbors=15). The expression relationships between cells were graphically visualized with the Python implementation [21] of the ForceAtlas2 [22] graph layout algorithm, as called in Scapny (‘tl.draw_graph’ and ‘pl.draw_graph’).

Annotation data regarding brain region, cortical layer, and GABAergic/glutamatergic/nonneuronal cell type features were extracted from Allen Brain Atlas sample data for mapping onto derived cell type clusters. The top 100 marker genes for each cell type+cortical layer (CT/CL) cluster (e.g., excitatory L4) were identified as

those with higher expression unique to each cluster by the Welch t test in SCANPY ('tl.rank_genes_groups'). The expression of the DB TF list genes, which were identified as marker genes in a CT/CL cluster, was mapped onto cluster figures.

Gene ontology analysis of target cluster marker genes was performed via the Protein Analysis Through Evolutionary Relationships (PANTHER) tool at the Geneontology.org web server [17]. Ontological terms associated with overabundance among cluster marker gene lists were established via Fisher's exact test, and the results were filtered by an FDR<0.05 (default setting); analyses were performed for biological process, molecular function, and cellular component terms. Disease term gene ontology analysis was performed via Enrichr [23,24], which is based on the ontology compiled by DisGeNET [25].

Results

High-scoring TFBSs differ between modern humans and Neanderthals and are enriched in homeobox and the brain

In the analysis of 21,990 SNPs identified in the comparison of the modern human and Neanderthal promoteromes—the collection of all human/Neanderthal proximal promoters of protein-coding genes—a total of 108 TF models, representing 110 unique differentially binding (DB) TF proteins—showed a significant difference in scoring (Wilcoxon rank statistical test) between the two hominin species (Table 1). Compared with the complete set of JASPAR TF genes from which they are derived, DB TF genes are strongly enriched among the homeobox genes (78/110; Fisher's exact test: odds ratio [OR] 12.40; p value 2.96×10^{-17}) and enriched in brain-specific expression (79/110; OR 2.86, p value 3.48×10^{-6}). Similarly, compared with all protein-coding genes, the complete set of TFs cataloged in the JASPAR database are themselves very strongly enriched for homeobox genes (OR 85.16; p value 5.79×10^{-177}) and moderately enriched for brain, eye, and retina genes (OR 1.80; p value 3.50×10^{-11}). DB TF genes were underenriched among immune-specific genes (29/110) compared with the JASPAR set (OR 0.37; p value 1.44×10^{-5}).

DB TFs enriched for vision, brain, and developmental terms

Enrichment analysis of the DB TF genes compared with the JASPAR TF genes was performed via the OmniPath annotation database [26,27] via Fisher's exact test, with Benjamini–Hochberg correction for

multiple testing and an FDR of 0.1. The DB TFs were over-enriched for 'Retina' (OR *inf*; FDR 0.015), 'Motor neurons' (OR 28.4; FDR 0.013), 'CUT homeobox' (OR 15.8; FDR 0.042), 'Vision' (OR 9.93; FDR 0.058), 'Sensory transduction' (OR 9.93; FDR 0.058), 'Bicoid' (OR 9.42; FDR 0.0078), 'Dwarfism' (OR 4.61; FDR 0.058), 'Developmental protein' (OR 2.89; FDR 1.56×10^{-4}), 'Brain' (OR 2.26; FDR 0.045), and 'Ectoderm' (OR 2.14; FDR 0.016), with the full results available in Supplementary Table 1.

Table 1. TFs with differential binding affinities in modern humans vs. Neanderthal proximal and semi-distal promoter regions. Dev., development-associated gene; HAR, human accelerated region.

Gene	Description	HAR	Dev.
ALX3	ALX Homeobox 3		x
ALX4	ALX Homeobox 4	x	x
ARID5A	AT-Rich Interaction Domain 5A		
BARHL1	BarH Like Homeobox 1		x
BARHL2	BarH Like Homeobox 2	x	x
BARX1	BARX Homeobox 1		x
BSX	Brain Specific Homeobox		x
CRX	Cone-Rod Homeobox		x
CUX1	Cut Like Homeobox 1	x	x
CUX2	Cut Like Homeobox 2		x
DLX1	Distal-Less Homeobox 1		x
DLX2	Distal-Less Homeobox 2	x	x

DLX4	Distal-Less Homeobox 4		x
DMBX1	Diencephalon/Mesencephalon Homeobox 1		x
DMRT3	Doublesex And Mab-3 Related Transcription Factor 3	x	
EMX1	Empty Spiracles Homeobox 1		x
EN1	Engrailed Homeobox 1	x	x
ESRRG	Estrogen Related Receptor Gamma	x	
ESX1	ESX Homeobox 1		x
EVX1	Even-Skipped Homeobox 1		x
EVX2	Even-Skipped Homeobox 2	x	x
FOS	Fos Proto-Oncogene, AP-1 Transcription Factor Subunit		
FOXA1	Forkhead Box A1	x	x
FOXD3	Forkhead Box D3		x
FOXK1	Forkhead Box K1		x
FOXK2	Forkhead Box K2		x
FOXP1	Forkhead Box P1	x	x
FOXP2	Forkhead Box P2	x	x
GBX2	Gastrulation Brain Homeobox 2	x	x
GFI1	Growth Factor Independent 1 Transcriptional Repressor		
GRHL2	Grainyhead Like Transcription Factor 2		

GSC	Goosecoid Homeobox		x
GSC2	Goosecoid Homeobox 2		x
GSX2	GS Homeobox 2		x
HAND1	Heart And Neural Crest Derivatives Expressed 1		
HMX1	H6 Family Homeobox 1		x
HMX2	H6 Family Homeobox 2		x
HMX3	H6 Family Homeobox 3		x
HNF1A	HNF1 Homeobox A		x
HNF1B	HNF1 Homeobox B		x
HOXA5	Homeobox A5	x	x
HOXB2	Homeobox B2		x
HOXB3	Homeobox B3		x
HOXB5	Homeobox B5		x
HOXD8	Homeobox D8		x
IRF3	Interferon Regulatory Factor 3		
ISL2	ISL LIM Homeobox 2		x
ISX	Intestine Specific Homeobox	x	x
JUN	Jun Proto-Oncogene, AP-1 Transcription Factor Subunit		
LBX2	Ladybird Homeobox 2		x

LEF1	Lymphoid Enhancer Binding Factor 1		
LHX3	LIM Homeobox 3		x
LHX4	LIM Homeobox 4		x
LHX9	LIM Homeobox 9		x
LMX1A	LIM Homeobox Transcription Factor 1 Alpha		x
LMX1B	LIM Homeobox Transcription Factor 1 Beta		x
MAFB	MAF BZIP Transcription Factor B	x	
MAFG	MAF BZIP Transcription Factor G		
MEF2C	Myocyte Enhancer Factor 2C	x	
MEOX1	Mesenchyme Homeobox 1		x
MEOX2	Mesenchyme Homeobox 2	x	x
MIXL1	Mix Paired-Like Homeobox		x
MNX1	Motor Neuron And Pancreas Homeobox 1		x
NEUROG1	Neurogenin 1		
NFYA	Nuclear Transcription Factor Y Subunit Alpha		
NKX2-5	NK2 Homeobox 5		x
NKX2-8	NK2 Homeobox 8		x
NKX6-2	NK6 Homeobox 2		x
NOBOX	NOBOX Oogenesis Homeobox		x

NR1H3	Nuclear Receptor Subfamily 1 Group H Member 3		
NR2E1	Nuclear Receptor Subfamily 2 Group E Member 1		
NRL	Neural Retina Leucine Zipper		
ONECUT1	One Cut Homeobox 1	x	x
ONECUT3	One Cut Homeobox 3		x
OTX1	Orthodenticle Homeobox 1		x
OTX2	Orthodenticle Homeobox 2		x
PAX3	Paired Box 3	x	x
PAX7	Paired Box 7		x
PHOX2A	Paired Like Homeobox 2A		x
PHOX2B	Paired Like Homeobox 2B	x	x
PITX3	Paired Like Homeodomain 3		x
POU1F1	POU Class 1 Homeobox 1	x	x
POU2F1	POU Class 2 Homeobox 1		x
POU2F3	POU Class 2 Homeobox 3		x
POU3F1	POU Class 3 Homeobox 1		x
POU3F2	POU Class 3 Homeobox 2	x	x
POU3F3	POU Class 3 Homeobox 3	x	x
POU4F1	POU Class 4 Homeobox 1		x

POU4F2	POU Class 4 Homeobox 2		x
POU4F3	POU Class 4 Homeobox 3	x	x
POU6F1	POU Class 6 Homeobox 1		x
POU6F2	POU Class 6 Homeobox 2	x	x
PROP1	PROP Paired-Like Homeobox 1		x
PRRX1	Paired Related Homeobox 1	x	x
PRRX2	Paired Related Homeobox 2		x
RARA	Retinoic Acid Receptor Alpha		
RAX	Retina And Anterior Neural Fold Homeobox		x
RAX2	Retina And Anterior Neural Fold Homeobox 2		x
RORA	RAR Related Orphan Receptor A	x	
RXRA	Retinoid X Receptor Alpha		
SHOX	Short Stature Homeobox		x
SOX10	SRY-Box Transcription Factor 10		
SOX15	SRY-Box Transcription Factor 15		
TBP	TATA-Box Binding		
TCF3	Transcription Factor 3		
UNCX	UNC Homeobox	x	x
VAX1	Ventral Anterior Homeobox 1		x

VAX2	Ventral Anterior Homeobox 2	x
VSX1	Visual System Homeobox 1	x
VSX2	Visual System Homeobox 2	x

DB TF target genes

A total of 74 genes were identified, hereafter referred to as 'DB TF target genes', whose promoters contain MH vs. Neanderthal SNPs where differential binding occurs by our determined set of statistically significant DB TFs (Table 2). In all of the DB TF target genes, at least one MH vs. Neanderthal SNP position has been identified as an eQTL affecting the expression of the target gene in whose promoter it occurs, as defined by data from the GTEx database (data not shown).

The target gene with the greatest number of DB events (143) was FARS2, whose promoter contains 2 SNPs, for which 105 (97.22%) of the DB TF binding models showed differential binding (Table 2). The remaining top ten DB TF target genes with the next highest number of DB occurrences were: TEX19, NARF, SLC6A11, LSM5, FAM172A, WNT7A, EXO1, KRT8, and ZNF37A. Among these genes, all but LSM5 have been associated with brain (FARS2, SLC6A11, FAM172A, WNT7A, and ZNF37A) or reproductive (FARS2, TEX19, NARF, WNT7A, EXO1, KRT8) tissue- or cell-specific expression in the Protein Atlas database (Supplementary Table 2). Among all 74 DB TF target genes, 30 have reproductive-specific expression, 28 have brain-specific expression, 18 have blood cell-specific expression, and 16 have mitochondria-specific expression or localization.

DB TF target genes and their direct interactors are enriched in the brain, physiology, vision, and autism

The OmniPath DB [26,27] was queried to identify all proteins that have posttranslational interactions with DB TF target proteins. This identified 805 genes that were combined with the 74 DB TF target genes for gene ontology enrichment analysis via the GeneOntology.org webserver [16,17]. Among the most enriched biological process terms associated with physiology were: 'midbrain morphogenesis' (GO:1904693; 22x enrichment), 'positive regulation of connective tissue replacement' (GO:1905205; 22x), 'trachea formation' (GO:0060440; 19x), 'regulation of midbrain dopaminergic neuron differentiation' (GO:1904956; 19x),

‘prostate gland morphogenetic growth’ (GO:0060737; 17x), ‘acinar cell differentiation’ (GO:0090425; 17x), ‘vitellogenesis’ (GO:0007296; 17x), ‘frontal suture morphogenesis’ (GO:0060364; 17x), ‘eyelid development in camera-type eye’ (GO:0061029; 15x), and ‘retinal blood vessel morphogenesis’ (GO:0061304; 15x). The complete biological process enrichment results are available in Supplementary Table 3.

Fisher’s exact test revealed enrichment of target genes and their direct interactors among autism spectrum disorder (ASD) genes from the SFARI gene knowledgebase (Banerjee-Basu and Packer 2010) (sfari.org; vers. 3.0), (OR 2.68; p value 1.16×10^{-16}).

Table 2. DB TF target genes with differential binding affinity by DB TFs in modern human vs. Neanderthal proximal and semi-distal promoter regions.

Gene	Gene description	Single-cell RNA cell specific	Count
FARS2	Phenylalanyl-tRNA synthetase 2, mitochondrial	astrocytes, early spermatids, excitatory neurons, inhibitory neurons, late spermatids, microglia, oligodendrocyte precursor cells, oligodendrocytes	143
LYRM4	LYR Motif-Containing Protein 4		143
TEX19	Testis expressed 19	spermatocytes, spermatogonia	60
NARF	Nuclear prelamin A recognition factor	erythroid cells, late spermatids	56
SLC6A11	Solute carrier family 6 member 11	astrocytes, excitatory neurons, inhibitory neurons, muller glia cells, oligodendrocyte precursor cells	54

LSM5	LSM5 homolog, U6 small nuclear RNA and mRNA degradation associated		49
FAM172A	Family with sequence similarity 172 member A	astrocytes, excitatory neurons, inhibitory neurons, microglia, oligodendrocyte precursor cells, oligodendrocytes	47
WNT7A	Wnt family member 7A	cytotrophoblasts, extravillous trophoblasts, pancreatic endocrine cells, syncytiotrophoblasts	43
EXO1	Exonuclease 1	early spermatids, erythroid cells, extravillous trophoblasts, plasma cells, spermatocytes, spermatogonia	42
KRT8	Keratin 8	distal enterocytes, ductal cells, exocrine glandular cells, extravillous trophoblasts, pancreatic endocrine cells, paneth cells, proximal enterocytes, syncytiotrophoblasts	35
ZNF37A	Zinc finger protein 37A		32
TIGD6	Tigger transposable element derived 6	spermatocytes	27
CSTF3	Cleavage stimulation factor subunit 3		21
GDPGP1	GDP-D-glucose phosphorylase 1		18
HLA-DQB1	Major histocompatibility	b-cells, langerhans cells, macrophages,	16

	complex, class II, DQ beta 1	monocytes	
LSM10	LSM10, U7 small nuclear RNA associated		15
ZNF132	Zinc finger protein 132	cardiomyocytes, cone photoreceptor cells, early spermatids	14
CCDC197	Coiled-coil domain containing 197		13
AK4	Adenylate kinase 4	hepatocytes, proximal tubular cells	13
DDX43	DEAD-box helicase 43	spermatocytes, spermatogonia	11
AKAP13	A-kinase anchoring protein 13		9
ZC3H3	Zinc finger CCCH-type containing 3		9
ZCCHC9	Zinc finger CCHC-type containing 9		9
MFAP3L	Microfibril associated protein 3 like	late spermatids	8
IFITM2	Interferon induced transmembrane protein 2	adipocytes, nk-cells	8
GPA33	Glycoprotein A33	distal enterocytes, intestinal goblet cells, paneth cells, proximal enterocytes, undifferentiated cells	8
MMEL1	Membrane	early spermatids, late spermatids, paneth cells	8

	metalloendopeptidase like 1		
PIGP	Phosphatidylinositol glycan anchor biosynthesis class P	spermatocytes	7
COA6	Cytochrome c oxidase assembly factor 6		7
PCNX4	Pecanex 4	dendritic cells, endometrial ciliated cells	7
PHGDH	Phosphoglycerate dehydrogenase	distal tubular cells, exocrine glandular cells	6
NPTXR	Neuronal pentraxin receptor	cardiomyocytes, excitatory neurons, inhibitory neurons, oligodendrocyte precursor cells	6
IMMT	Inner membrane mitochondrial protein	cardiomyocytes	6
CDC37L1	Cell division cycle 37 like 1		5
SMC2	Structural maintenance of chromosomes 2	endometrial ciliated cells	5
AGR2	Anterior gradient 2, protein disulphide isomerase family member	club cells, distal enterocytes, gastric mucus-secreting cells, intestinal goblet cells, paneth cells, undifferentiated cells, urothelial cells	4
SLC8A3	Solute carrier family 8 member A3	bipolar cells, excitatory neurons, inhibitory neurons, oligodendrocyte precursor cells	4
AKT1S1	AKT1 substrate 1	syncytiotrophoblasts	4
HNRNPUL1	Heterogeneous nuclear		4

	ribonucleoprotein U like 1		
RNASEH2A	Ribonuclease H2 subunit A	cytotrophoblasts, extravillous trophoblasts	3
PRR18	Proline rich 18	oligodendrocytes	2
KCNQ3	Potassium voltage-gated channel subfamily Q member 3	excitatory neurons, inhibitory neurons, microglia	2
ITGA1	Integrin subunit alpha 1	adipocytes, endothelial cells, hepatocytes, sertoli cells, smooth muscle cells	2
GRIN2A	Glutamate ionotropic receptor NMDA type subunit 2A	excitatory neurons, inhibitory neurons	2
TACC3	Transforming acidic coiled-coil containing protein 3	early spermatids, hofbauer cells, late spermatids, spermatocytes	2
EIF2A	Eukaryotic translation initiation factor 2A		2
MPHOSPH9	M-phase phosphoprotein 9	cone photoreceptor cells, excitatory neurons, inhibitory neurons	2
F7	Coagulation factor VII	hepatocytes	2
ZNF211	Zinc finger protein 211		2
KLHL24	Kelch like family member 24	oligodendrocytes	2
DKKL1	Dickkopf like acrosomal protein 1	early spermatids, late spermatids	2

STPG2	Sperm tail PG-rich repeat containing 2	astrocytes, excitatory neurons, inhibitory neurons, microglia, oligodendrocyte precursor cells, oligodendrocytes	1
PLD3	Phospholipase D family member 3	hofbauer cells	1
MZT1	Mitotic spindle organizing protein 1	erythroid cells	1
EVI5	Ecotropic viral integration site 5	oligodendrocytes	1
DGCR2	DiGeorge syndrome critical region gene 2		1
GABRA5	Gamma-aminobutyric acid type A receptor subunit alpha5	excitatory neurons, inhibitory neurons	1
CYP4F12	Cytochrome P450 family 4 subfamily F member 12	distal enterocytes, gastric mucus-secreting cells, paneth cells, proximal enterocytes	1
RNF215	Ring finger protein 215	bipolar cells, cone photoreceptor cells, rod photoreceptor cells	1
ADARB2	Adenosine deaminase RNA specific B2 (inactive)	inhibitory neurons, oligodendrocyte precursor cells	1
MYL12A	Myosin light chain 12A	cardiomyocytes	1
RUVBL2	RuvB like AAA ATPase 2	early spermatids, late spermatids, respiratory ciliated cells, spermatocytes	1

TRIM2	Tripartite motif containing 2	oligodendrocytes	1
CADPS	Calcium dependent secretion activator	bipolar cells, cone photoreceptor cells, excitatory neurons, inhibitory neurons, oligodendrocyte precursor cells	1
PCK1	Phosphoenolpyruvate carboxykinase 1	distal enterocytes, distal tubular cells, hepatocytes, proximal enterocytes, proximal tubular cells	1
NGDN	Neuroguidin		1
ABCA8	ATP binding cassette subfamily A member 8	fibroblasts, granulosa cells, leydig cells, microglia, oligodendrocytes, skeletal myocytes	1
POGZ	Pogo transposable element derived with ZNF domain	oligodendrocytes	1
PDIA6	Protein disulfide isomerase family A member 6	extravillous trophoblasts, plasma cells	1
ACSL1	Acyl-CoA synthetase long chain family member 1	early spermatids, hepatocytes, late spermatids	1
SUPV3L1	Suv3 like RNA helicase		1
DCXR	Dicarbonyl and L-xylulose reductase	distal tubular cells, hepatocytes, proximal tubular cells	1
DPYSL5	Dihydropyrimidinase like 5	cone photoreceptor cells, late spermatids, oligodendrocytes	1
OMA1	OMA1 zinc metallopeptidase	oligodendrocytes	1

NME1	NME/NM23 nucleoside diphosphate kinase 1	1

DB TFs are coexpressed in neural and immune tissues

Data derived from the bulk RNA-Seq experiments (312 different tissues; 686 samples) of the FANTOM5 project were used to generate a cluster map of DB TF expression within the 100 tissues with the highest aggregate expression of the DB TFs, presented in Figure 5A (y-axis DB TFs; x-axis, samples). On the basis of their tissue expression in the FANTOM dataset, the DB TFs form two distinct top-level clusters: the larger cluster contains 89 genes, the majority of which are known to have a role in development (78/89) and specific expression in brain or retina tissue (72/89), whereas the smaller cluster of 17 genes all show specific expression in blood and immune cells. Data on tissue group specificity (brain and retina; male/female) from the Human Protein Atlas (bulk and single-cell RNA-Seq), status as a development-related gene (homeobox, forkhead box, and SRY-related HMG-box), and above the 90th quantile expression in immune cell types as defined by the Database of Immune Cell eQTLs (DICE), were analyzed and included in Figure 5C.

Within the larger cluster of 89 developmental- and brain-specific genes, there is a subcluster of 12 genes whose expression is most limited to FANTOM 5 neural tissues: NKX6-2, CUX2, NR2E1, POU3F2, POU6F2, EMX1, VAX1, POU3F3, DLX1, DLX2, OTX1, and POU3F1. All of these genes are identified in the Protein Atlas as being enriched or enhanced in the brain or retina. Compared with all protein-coding genes, GO analysis of these 12 genes with the PANTHER-based gene ontology.org web server [16,17] revealed the top 10 enriched biological process terms: ‘cerebral cortex GABAergic interneuron fate commitment’ (GO:0021893; 100x enrichment), 'positive regulation of amacrine cell differentiation' (GO:1902871; 100x), 'forebrain ventricular zone progenitor cell division' (GO:0021869; 100x), 'negative regulation of photoreceptor cell differentiation' (GO:0046533; 100x), 'regulation of photoreceptor cell differentiation' (GO:0046532; 100x), 'regulation of amacrine cell differentiation' (GO:1902869; 100x), 'neuroblast differentiation' (GO:0014016; 100x), 'negative regulation of oligodendrocyte differentiation' (GO:0048715; 100x), 'cerebral cortex GABAergic interneuron

differentiation' (GO:0021892; 100x), and 'negative regulation of glial cell differentiation' (GO:0045686; 100x).

Additionally, within the brain tissue cluster, we observed subclusters of fetal and newborn tissues (Figure 2D). Overall, DB TFs presented the highest expression in the medial frontal gyrus (newborn), medial temporal gyrus (adult), parietal lobe (fetal), occipital lobe (fetal), eye (fetal), pineal gland (adult), temporal lobe (fetal), retina (adult), occipital cortex (newborn), parietal lobe (newborn), dura mater (adult), medial temporal gyrus (newborn), and spinal cord (fetal) (Supplementary Figure 1). Similarly, the cluster with high expression in blood and immune cells/tissues is comprised of 17 genes: FOS, RORA, FOXK1, FOXK2, IRF3, NFYA, TBP, MAFG, POU2F1, ARID5A, RARA, JUN, CUX1, FOXP1, MEF2C, MAFB, and RXRA. All of these genes have an expression ≥90th quantile (compared with all protein-coding genes) in at least one immune cell type, as cataloged in the Database of Immune Cell eQTLs (DICE) [28]. GO analysis of the 17 genes (geneontology.org web server) revealed enrichment of immune-related biological process terms: ‘CD4-positive, alpha-beta T-cell differentiation involved in immune response’ (GO:0002294; 76x enrichment), ‘T-helper cell differentiation’ (GO:0042093; 76x), and ‘alpha-beta T-cell differentiation involved in immune response’ (GO:0002293; 73x), among others.

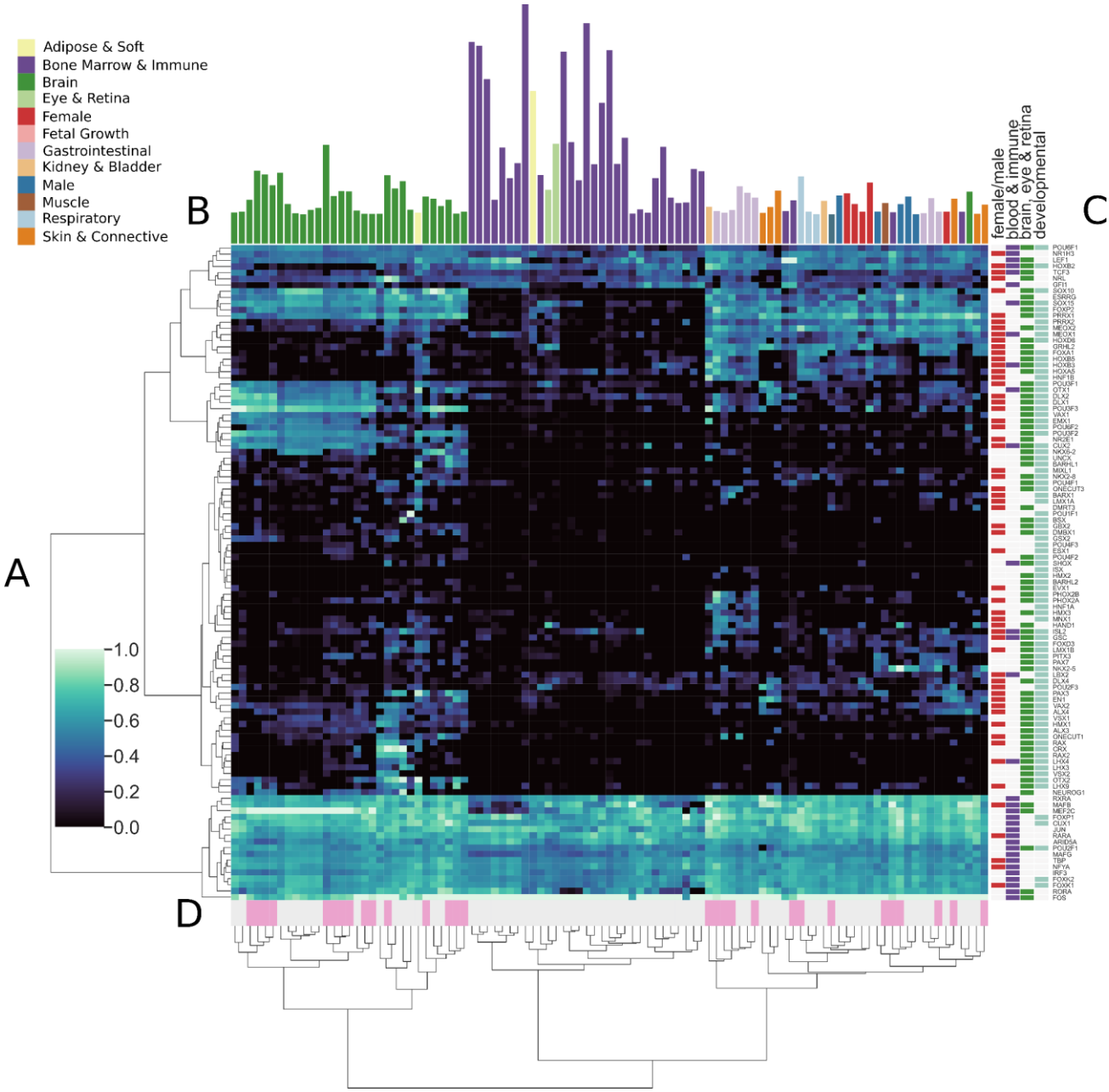


Figure 2. Cluster map of the expression of transcription factors displaying variable binding in modern human vs. Neanderthal. (**A**) Expression data derived from the FANTOM 5 dataset were used to cluster 106 DB transcription factors (y-axis) in 100 tissues (x-axis). Expression data were not available for the DUXA and PROP1 genes. (**B**) Aggregate expression for each tissue across all genes is presented as bars, and bars/tissues are colored by tissue group. Distinct clusters of brain and eye/retina tissues as well as blood and immune cells/tissues are observed. A bar plot with the tissue names is presented in Supplementary Figure 1. (**C**) TF genes are labeled with colored bars corresponding to their identification as homeobox, SOX, or FOX genes (developmental), above the 90th quantile expression in an immune cell type as defined by the DICE database

(blood & immune) [28], or to have enriched or enhanced expression by tissue as defined by proteinatlas.org (female/male; brain, eye & retina) [29]. (**D**) Life stage for each tissue is described, where white and pink represent adult and fetal/newborn tissues, respectively. The expression values (TPM) were normalized via log transformation, and each tissue column was adjusted on a zero-to-one scale.

DB TFs have developmental time-point-related expression profiles in the brain

Cluster analysis of the Allen Brain Atlas bulk RNA-Seq expression data for DB TF genes revealed distinct clusters of time-point-specific expression. Those genes comprising the cluster with the highest aggregate expression in the dataset (module 1 genes: NFYA, TCF3, MEF2C, SOX10, CUX2, TBP, FOXP1, IRF3, SOX15, RARA, FOXK2, NKX6-2, FOXK1, FOS, JUN, POU6F1, POU3F2, MAFG, POU3F3, MAFB, and CUX1) show a dynamic and cyclic pattern of expression during the prebirth timepoints, which then stabilizes in infancy and trends downward in early childhood (Figure 3A). Despite differences in physiology, different tissues generally cluster together by time point rather than tissue type: pcw (8–37 weeks post conception), early (4 months–4 years), and late (8 years–40 years) (Figure 3B).

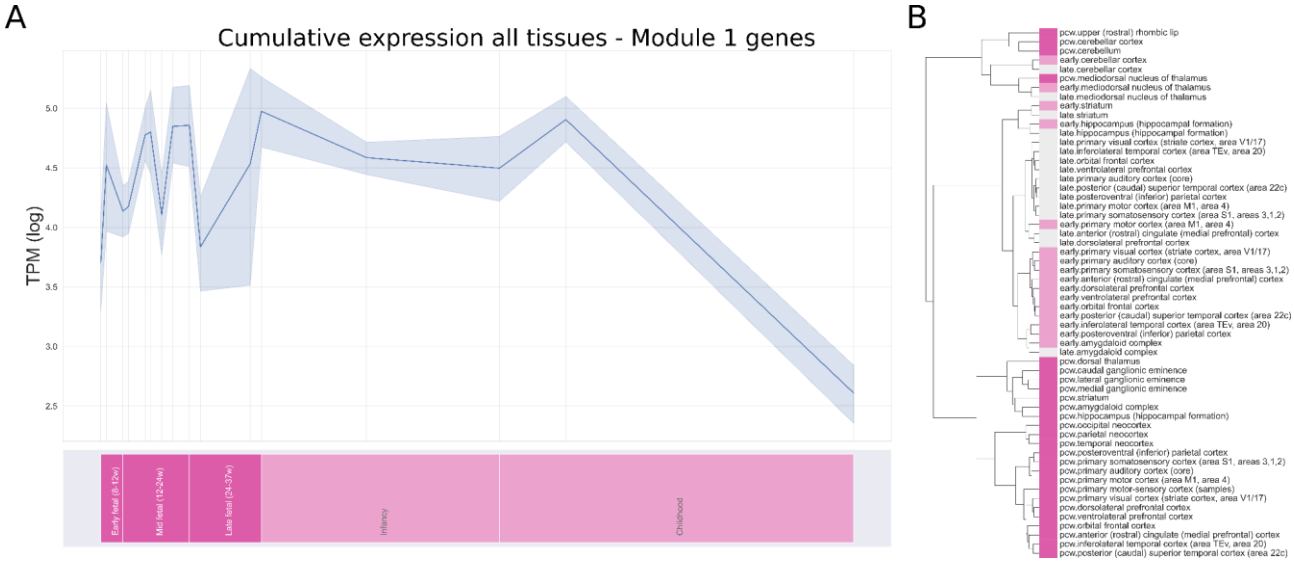


Figure 3. Expression of modern human vs Neanderthal DB TFs in developing brain tissue. The expression data from the Allen Brain Atlas [30] were extracted as RPKM values for 26 unique tissues across 31 timepoints, converted to TPM values, log transformed, and used to construct a cluster map. A) Aggregate expression of the cluster of DB TF genes with the highest total expression across all tissues (module 1) from 8–144 weeks post conception. B) Clustering of brain tissues defined by development time point. The colors represent development stages for both figures: 8–37 weeks post conception (pcw, magenta), 4 months–4 years (early, pink), and 8–40 years (late, white).

Single-nucleus RNA-Seq of cortical cells reveals differential expression of DB TFs

An analysis of the Allen Brain Atlas single-nucleus RNA-Seq data from 49,417 cortical brain cells [18] was performed to identify which of the MH-Neanderthal DB TF genes may play a functional role in specific annotated brain regions, layers, and cell types. The analyses in the original paper defined metadata indicating cell class (excitatory, inhibitory, and nonneuronal), layer (L1–L6), and predicted type (e.g., astrocyte, microglia, endothelial, etc.), as well as brain regions (Figure 4A). Cell groups defined by the intersection of class, layer, and region were analyzed for differentially expressed genes (DEGs). Several DB TFs were DEGs in specific categories, occurring in nearly all cases in glutamatergic neurons, usually in layers L4–L6, and across nearly all brain regions. The categories in which four or more DB TFs were DEGs are presented in Table 3 and included cut-like homeobox 1 (CUX1), cut-like homeobox 2 (CUX2), estrogen-related receptor

gamma (ESRRG), forkhead box protein 1 (FOXP1), forkhead box protein 2 (FOXP2), Myocyte Enhancer Factor 2C (MEF2C), POU Class 6 Homeobox 2 (POU6F2), paired related homeobox 1 (PRRX1), and RAR-related orphan receptor alpha (RORA). In all categories, POU6F2 was a DEG, and in all but two categories, either FOXP2 or FOXP1 (or both) were DEGs. FOXP2 was present as a DEG in excitatory L4 cells in A1C (primary visual cortex), MTG (medial temporal gyrus), S1lm (lower limb somatosensory cortex), and S1 µl (upper limb somatosensory cortex).

Table 3. DB TFs that co-occur as DEGs in specific regions, layers, and cell types according to the Allen Brain Atlas scRNA-Seq data from the cerebral cortex

Layer	Class	Region	DEG DB TF genes
L4	Glutamatergic	A1C	FOXP2, FOXP1, POU6F2, MEF2C
L4	Glutamatergic	MTG	FOXP2, PRRX1, POU6F2, MEF2C
L4	Glutamatergic	S1lm	FOXP2, PRRX1, POU6F2, ESRRG
L4	Glutamatergic	S1ul	FOXP2, FOXP1, RORA, POU6F2, MEF2C
L4	Glutamatergic		FOXP1, CUX2, RORA, POU6F2, CUX1, MEF2C
L4		IT	FOXP1, CUX2, RORA, POU6F2, CUX1, MEF2C
L4ab	Glutamatergic	V1C	CUX2, RORA, POU6F2, CUX1, MEF2C
L4ab			CUX2, RORA, POU6F2, CUX1
L5	Glutamatergic	S1lm	FOXP2, PRRX1, POU6F2, ESRRG
L5	Glutamatergic	V1C	FOXP2, FOXP1, RORA, POU6F2, MEF2C
L6a			FOXP2, PRRX1, POU6F2, ESRRG

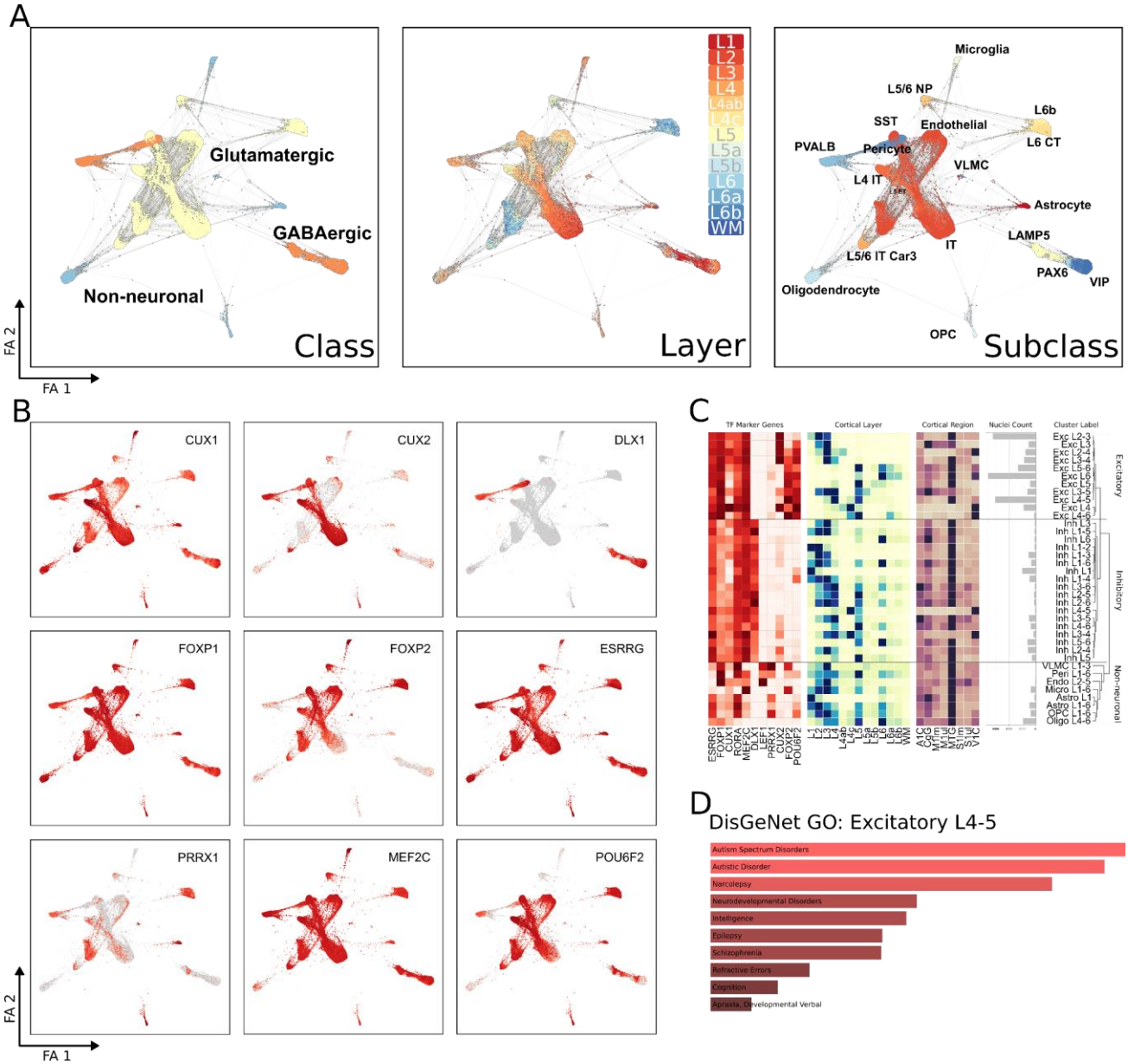


Figure 4. Analysis of single-nucleus RNA-Seq data from brain cortical regions. (**A**) Force-directed graph (FA) visualization of Allen Brain Atlas single-nucleus RNA-Seq data from 49,417 brain cortex cells; the brain excitatory/inhibitory status, cortical layer, and predicted cell type (subclass) data are mapped from annotations in the Allen Brain Atlas sample data [18]. (B) Expression of DB TFs, which are DEGs mapped onto FA visualization. C) On the x-axis: DB TF DEGs, cortical layer, and cortical region; on the y-axis: cell class/type and cortical layer, with the dendrogram indicating expression relationships. The color intensity indicates the normalized expression (0–1 scale). Nuclei count indicates the number of nuclei for that cell class/type at the indicated layer. D) DisGeNET disease ontology terms enriched for the top 100 marker genes for L4–5 excitatory cells. ET, extratelencephalic/pyramidal tract; IT, intratelencephalic; NP, near-projecting; OPC,

oligodendrocyte precursor; VLMC, vascular leptomeningeal cell; SST, LAMP5, PAX6, VIP, PVALB, labels on the basis of expression of the corresponding gene.

Gene ontology analysis of L4 excitatory neurons

In the Allen Brain Atlas cortical cell snRNA-Seq data, multiple DB TFs were consistently up-regulated DEGs in both layer 4 and excitatory cell groups, and their intersection. Gene ontology analysis of the top 100 marker genes of L4 excitatory cells via the DisGeNet ontology within the Enrichr tool [24] revealed the top enriched terms ‘autism spectrum disorders’, ‘autistic disorder’, ‘narcolepsy’, ‘neurodevelopmental disorders’, ‘intelligence’, ‘epilepsy’, ‘schizophrenia’, ‘refractive errors’, ‘cognition’, and ‘apraxia, developmental verbal’ (Figure 4D). The top 10 enriched biological process terms for L4 excitatory cells were: ‘corticospinal neuron axon guidance’ (GO:0021966; >100x enrichment), ‘corticospinal tract morphogenesis’ (GO:0021957; 91x), ‘ventricular cardiac muscle cell differentiation’ (GO:0055012; 42x), ‘synaptic membrane adhesion’ (GO:0099560; 35x), 'central nervous system projection neuron axonogenesis' (GO:0021952: 30x), 'synaptic transmission', GABAergic' (GO:0051932: 29x), 'positive regulation of excitatory postsynaptic potential' (GO:2000463: 28x), 'negative regulation of smooth muscle cell migration' (GO:0014912: 27x), 'positive regulation of dendrite morphogenesis' (GO:0050775: 25x), and 'modulation of excitatory postsynaptic potential' (GO:0098815: 25x).

Gene ontology analysis of L4 excitatory neurons

In the GO analysis of the CC category, the top 10 enriched terms for L4 excitatory cells were: ‘anchored component of presynaptic membrane' (GO:0099026; 61x enrichment), ‘presynaptic cytosol' (GO:0099523; 47x), ‘NMDA selective glutamate receptor complex' (GO:0017146; 47x), ‘intrinsic component of presynaptic active zone membrane' (GO:0098945; 45x), ‘node of Ranvier' (GO:0033268; 42x), ‘integral component of presynaptic active zone membrane' (GO:0099059; 40x), ‘GABA-A receptor complex' (GO:1902711; 34x), ‘GABA receptor complex' (GO:1902710; 30x), ‘presynaptic active zone membrane' (GO:0048787; 27x), and ‘intrinsic component of presynaptic membrane' (GO:0098889; 22x).

In the GO analysis of molecular function, the top 10 enriched terms for L4 excitatory cells were: 'transmembrane receptor protein tyrosine phosphatase activity' (GO:0005001; 50x enrichment),

'transmembrane receptor protein phosphatase activity' (GO:0019198; 50x), 'GABA-gated chloride ion channel activity' (GO:0022851; 49x), 'inhibitory extracellular ligand-gated ion channel activity' (GO:0005237; 42x), 'ligand-gated anion channel activity' (GO:0099095; 35x), 'GABA-A receptor activity' (GO:0004890; 34x), 'syntaxin-1 binding' (GO:0017075; 32x), 'GABA receptor activity' (GO:0016917; 29x), 'transmitter-gated channel activity' (GO:0022835; 17x), and 'transmitter-gated ion channel activity' (GO:0022824; 17x).

DB TFs and Human Accelerated Regions

Doan et al. identified and cataloged a list of genes in close proximity to human accelerated regions (HARs; elevated divergence in humans vs. other primates) as well as those that directly interact with HARs through chromatin conformation changes, e.g., enhancers [31]. Among the set of 110 DB TF genes, 28 were HAR-associated. Fisher's exact test revealed that the overrepresentation of HAR associations in DB TFs versus all JASPAR TFs was not significant (OR 0.98; p value 0.57). However, compared with all protein-coding genes, the JASPAR TFs as a group were enriched for HAR associations (OR 3.27; p value 6.73×10^{-26}).

Discussion

Regulatory changes drive evolution

Current knowledge indicates that, across the tree of life, prior to and during major radiations of new species, there is a commensurate increase in the number of regulatory genes. Comparative genomic analyses suggest that at the time of eukaryogenesis, or the origin of the last eukaryotic ancestor (LECA), a significant increase in novel TF classes occurred [32]. Eukaryogenesis is one of the major transitions of life on Earth, with increased diversity owing to the endosymbiotic synthesis of an energy-producing α-proteobacteria mitochondrion-progenitor within an archaeon [33]. If the new in-house energy source could be described as the engine driving the diversity arising from LECA, the complementary argument could be that it was TF proteins that performed the steering. Later, prior to the colonization of land by plants, there was an increase in the TF families of the ancestral aquatic streptophytes, which were then already present in the first land plants, which benefited from rapid diversification thereafter in the new environment [34]. Prior to the radiation of bilaterian (multicellular) metazoans, an increase in transcription factor families occurred [35], roughly

quadrupling in ratio [36]. Arguably, these increases in TF family numbers immediately prior to major transitions in life could indicate the critical role of TFs in the rise and adaptation of eukaryotes and the complexity arising within their successors.

Far more recently, a transposon-mediated shift in regulatory signaling in mammalian pregnancy has led to the endometrial stromal cell type [37] and the rewiring of a stress response producing the decidual stromal cell type [38]. In humans, the development of our cognitive skills is possibly due to a delay in synaptic-function gene expression and corresponding synaptogenesis in the prefrontal cortex, owing to transcription factors such as myocyte enhancer factor 2A (MEF2A) [39–42]. At the same time, evidence is accumulating that maturation occurred earlier in the closely related Neanderthal, as it does in chimpanzees [43], which is supported by samples of jaw [44], tooth [45], and, most importantly, cranium [46]. Furthermore, the development of a globular brain structure was not present in modern humans at the time of divergence from the Neanderthal and Denisovan lineages but has been a unique product of modern human development within the last 35,000–100,000 years [47], making it particularly attractive for further regulatory analyses.

Modern human brain regulatory changes

Compared with the complete set of JASPAR TFs, which are very strongly enriched compared with all protein-coding genes, the TF genes identified as differentially binding genes are strongly enriched for homeobox genes, the primary regulators of development in humans and other multicellular organisms. Similarly, the DB TFs are enriched for brain-specific expression compared with the JASPAR TFs, which are again enriched compared with all human genes. Hierarchical clustering of TF expression revealed that DB TF genes associated with neural and immune functions formed distinct groups. Within the tissue axis, in the neural cluster, adult and fetal/newborn neural tissues cluster separately in several instances (Figure 2). From these discoveries in a broader set of data, it became important to focus more closely on the expression of these DB TFs in developing neural tissues specifically. Analysis of RNA-Seq data from whole-brain subtissues further revealed differences in the expression of these genes across time and likely in a developmentally dependent manner. Specifically, the expression of DB TFs in all tissues for the two groups occurring within 8 weeks post conception to 4 years post birth was largely segregated from all tissues for the group defined by 8 years to 40 years of age (Figure

3). Additionally, several modules of DB TFs identified by cluster analysis appear to show a cyclical pattern of expression during fetal development, which then stabilizes post birth (Figure 3A).

DB TFs co-occur as DEGs according to single-cell RNA-Seq of glutamatergic cells in the L4 human cortex

Recently, the proliferation of single-cell transcriptomics has begun to clarify the direct role of transcription factors in cell type determination and identity (Arendt et al. 2019). To determine in which cell types DB TFs are expressed in the adult brain, single-nucleus data for 49,417 cortical cells were examined. A subset of the DB TFs were annotated as commonly co-occurring DEGs: CUX1, CUX2, ESRRG, FOXP1, FOXP2, MEF2C, POU6F2, PRRX1, and RORA. Interestingly, we observed that these DB TF DEGs were primarily co-occurring in glutamatergic cells of the L4 cortex. Among this set of nine genes, six (CUX1, CUX2, FOXP1, FOXP2, MEF2C, and RORA) are cataloged in the SFARI Gene knowledgebase [48] (sfari.org; vers. 3.0) as ASD candidate genes. Similarly, six (CUX1, FOXP1, FOXP2, MEF2C, POU6F2, and RORA) are present in the genome-wide association studies (GWAS) catalog [49] (www.ebi.ac.uk/gwas) for the 'schizophrenia' (MONDO_0005090) term. In support of this connection, a recent large-scale study on schizophrenia genetics in 74,776 individuals with schizophrenia (and 101,023 control individuals) revealed enrichment of genes that are highly expressed in glutamatergic neurons of both the human and mouse cerebral cortex [50]. Similarly, a recent large-scale study on ASD genetics in 11,986 individuals with ASD (and 23,598 control individuals) identified 102 ASD risk genes and revealed that fetal-development excitatory neurons accounted for the greatest number of these genes among all the identified cell types [51].

Co-occurring DB TF DEGs associate with HARs and neuropsychiatric disease

Human-accelerated regions (HARs) are locations in the genome where the rate of evolutionary change has accelerated since divergence from chimpanzee. A study of 2,737 HARs revealed that they are enriched for TFBSs generally and for TFs associated with neural development specifically (Doan et al. 2016). While the DB TFs were not enriched for HAR association compared with the JASPAR TFs (OR 0.98; p value 0.57), the JASPAR TFs are compared to all genes (OR 3.27; p value 6.73×10^{-26}). Using the HARs cataloged by [31], we identified 28 DB TFs that are either in close proximity to a HAR or directly interact with one through chromatin conformation changes. Additionally, all but one (CUX2) of the co-occurring DEGs DB TFs (CUX1,

ESRRG, FOXP1, FOXP2, MEF2C, POU6F2, PRRX1, and RORA) identified in the L4 excitatory neurons of the cerebral cortex (Allen Brain Atlas scRNA-Seq data) contain HAR regions or HAR interactivity. Several of these contain or associate with HAR variants that have been noted to cause neurological disease. POU6F2 possesses an intronic HAR in which a mutant allele (GRCH38:chr7:39,033,595) is associated with ASD (Doan et al. 2016). The promoter of the CUX1 gene has been determined via ChIA-Pet analysis of chromatin to interact with a HAR located ~200 kb away; unrelated individuals with intellectual disability (ID) (IQ<40) and ASD have been identified with a homozygous mutation (GRCh38:chr7:101,606,361) in this HAR. A luciferase reporter assay revealed that when the mutant version of this HAR interacts with the promoter of the CUX1 gene, its expression is increased threefold, while cultured differentiated neurons with increased CUX1 expression exhibit increased synaptic spine density. Similarly, another HAR (GRCh38:chr5:88,480,873) has been shown to interact with the promoter of MEF2C, and a mutation in this HAR creates a putative MEF2A binding site, reducing expression by ~50%. Mutations in the MEF2C gene are associated with autism [52,53], mental retardation [52–54], schizophrenia [55,56], epilepsy [52,53], and speech abnormalities [52]. Additionally, binding motifs for several of the DB TFs identified in our study were found to be enriched in HAR regions; POU6F1 and POU2F1 in ultra-conserved HARs, and HNF1A in all HARs [31]. After these intriguing results, further direct analysis of HARs via the TFBSFootprinter tool is warranted.

Importantly, one of the identified DB TFs in our data is the FOXP2 TF gene, which was the first noted "speech gene" [57]. For the early postnatal period in mice, FOXP2 is a negative regulator of MEF2C, and the likely result is the promotion of synaptogenesis in the cortical striatum [58]. In addition to this interaction with MEF2C, there is a POU3F2 (which, according to our results, is also a DB TF) binding site within the FOXP2 gene, which has been shown to affect FOXP2 expression and is associated with a selective sweep that has occurred since the divergence of humans and Neanderthal [59,60]. Because of its significant role in human evolution, further analysis is warranted regarding the regulation of FOXP2 in other relevant areas (e.g., introns, 3' UTRs, enhancers, etc.) and for genome-wide binding sites of FOXP2 itself.

DB TF target genes are expressed in oligodendrocytes, testes/sperm, the maternal–fetal interface, and mitochondria

There were 74 genes identified as DB TF target genes. All but 8 of these genes are specifically expressed in reproductive (30), brain and retina (28), immune (18), or mitochondrial (16) cells or tissues. Among the 30 DB TF target genes with reproductive cell- and tissue-specific expression, the majority (19) are expressed in male tissues (testis, early and late spermatids, spermatocytes, and spermatogonia). A total of 12 genes were expressed in female-specific cells and tissues and, in many cases, specifically in those involved in the maternal–fetal interface (cervix, granulosa cells, cytotrophoblasts, extravillous trophoblasts, syncytiotrophoblasts, and endometrial ciliated cells). Among the 28 DB TF target genes with brain- and retina-specific expression, 16 are expressed in oligodendrocytes, 12 in inhibitory neurons, and 11 in excitatory neurons.

The DB TF target gene with the greatest number of differential binding events in its promoter region is FARS2, a member of the mitochondrial aminoacyl-tRNA synthetases (mt-aaRSs). scRNA-Seq data from the Protein Atlas indicate that FARS2 expression is increased in excitatory neurons, oligodendrocyte precursor cells, inhibitory neurons, oligodendrocytes, astrocytes, microglia, late spermatids, and early spermatids. mt-aaRSs charge their cognate mt-tRNA with the appropriate amino acid, and mutations in these genes cause a variety of diseases, most predominantly those that affect the central nervous system [61], perhaps due to delayed myelination, demyelination, or both [62]. In addition to their canonical roles, mt-aaRSs are hypothesized to function in monitoring amino acid levels, as sensors for the mitochondrial environment, and in transcriptional regulation [61], and through the addition of new protein domains have been associated with neural development and immune response, among other processes, as reviewed by [63]. The FARS2 gene produces the mt-PheRS protein (phenylalanyl-tRNA synthetase), which is mitochondria-locating and is responsible for attaching phenylalanine to its corresponding mt-tRNA for mitochondrial protein translation [64]. Intragenic variants in the FARS2 gene have been linked to two primary clinical manifestations, early-onset epileptic mitochondrial encephalopathy and spastic paraplegia, and for patients in both groups, symptoms can also include intellectual disability or developmental delay [64]. Deletions within FARS2 and reduced expression levels have also been associated with schizophrenia [65].

Importantly, FARS2 shares a bidirectional promoter with another mitochondrial gene, LYRM4 (LYR motif-containing protein 4), which has a TSS just ~400 bp away. Together with ISCU, NFS1, and FXN, the ISD11 protein of the LYRM4 gene is involved in the formation of iron–sulfur (Fe–S) clusters [66,67], which are essential cofactors in many basic biological processes, such as the formation of respiratory chain complexes I, II, and III [68] and subsequent oxidative phosphorylation [69]. Mutations in LYRM4 cause deficits in oxidation phosphorylation reactions [70], which are critical in neuronal development and schizophrenia, as reviewed previously [71]. In support of this connection, polymorphisms occurring in the FARS2-LYRM4 bidirectional promoter region have been shown to be associated with cognitive deficit and schizophrenia [72]. Similarly, a 900 kb microdeletion in this region (encompassing RPP40, PPP1R3G, LYRM4, and part of the FARS2 and CDYL genes) has been shown to produce gyral pattern anomalies, intellectual disability and speech and language disorders [73].

A comparison of metabolite levels in the prefrontal cortex, visual cortex, cerebellum, kidney, and muscle between humans and other primates (chimpanzee and macaque), showed ‘aminoacyl-tRNA biosynthesis’ was the most significantly enriched pathway for metabolites having higher levels in human; for all three brain regions, but not muscle or kidney [74]. Similarly, the ‘Phenylalanine, tyrosine and tryptophan biosynthesis’ pathway was enriched for all three brain regions. This observation fits with the ongoing hypothesis that there is a requirement for the coevolution of nuclear and mitochondrial genes coding for mitochondrial proteins, known as mitonuclear compensatory coevolution, as reviewed in [75], and that this coevolution is a driver of speciation [76,77]. In the mitonuclear compensatory coevolution hypothesis, nuclear-encoded genes encoding aminoacyl tRNA synthetases (such as FARS2) and OXPHOS complex components (such as LYRM4) are expected and, in some cases, have been shown to evolve more rapidly in response to high rates of evolutionary change in mitochondrially encoded genes, with which their protein products subsequently directly interact [75]. Another consideration is how to regulate the translation of OXPHOS complex components when the corresponding genes exist in both the mitochondrial and nuclear genomes [78,79]. The high number of observed DB TFBSs in the bidirectional promoter of the nuclear FARS2 and LYRM4 genes thus potentially allows more well-regulated coordination of cross-compartment OXPHOS component gene expression and may even contribute to the effects of both mitonuclear compensation and speciation.

Taken together, the LYRM4-FARS2 locus is potentially of great interest in the divergence of modern humans from other hominins and hominids. It is reasonable to imagine that differences in neuronal development and function, and even diet, between MH and Neanderthal could require corresponding differences in mitochondrial activity, metabolism, and amino acid sensing.

Lots of work left to do

The divergence of modern humans and Neanderthals is estimated to have occurred between 400,000 and 800,000 years ago [80–82]. Since then, both subspecies have experienced unique evolutionary, social, and cultural paths, many of which overlap and intertwine. What appears to be both divergent and unifying in the hominin lineage is a significant change in the brain, leading to differing abilities in terms of cognition, social function, language and creativity. In the case of humans, and likely for Neanderthals as well, many of the same genes, which are at the core of our novel capabilities, are the same which are commonly the root of our maladies. Among other neuropsychiatric disorders, autism and schizophrenia appear to be maintained in humans at a greater rate than would be expected if they were strictly deleterious: the global prevalence of ASD is approximately 1% [83], and 0.28% for schizophrenia [84]. Likewise, as the study of these diseases continues, the direction of understanding tends toward a continuum of effects rather than a discrete on/off state. A study by Linscott et al. reported that the prevalence of psychotic experiences in the general population was 7.2% [85]. Our results revealed that DB TFs are enriched for development and the brain, and studies on HAR regions have shown that mutations in regulatory regions alone are sufficient to cause significant neuropsychiatric effects [31]. As a result, there appear to be significant opportunities to analyze gene regulation to understand not only the evolution of human cognitive abilities but also the neuropsychiatric disorders that accompany them.

Limitations

We have not included analyses of 3' UTRs, introns, or enhancers. The single-most expressed transcript for each protein-coding gene was chosen for analysis, while MH vs. Neanderthal SNPs may have occurred more closely to TSSs of lower-expressed transcripts and thereby had higher and perhaps more significant scoring. Only transcripts for protein-coding genes were analyzed, despite the existence of numerous other transcribed sequence types. The SNP dataset we used for analysis is based on multiple Neanderthal individuals and has

lower sequencing coverage than newer datasets. The inclusion of multiple individuals is useful for comparing modern humans to Neanderthals as a group but can provide only a more general comparison. Using a higher coverage dataset would allow greater assurance that the SNPs under inquiry are legitimate.

Acknowledgments

The nonprofit CSC—IT Center for Science Ltd., owned by the state of Finland and Finnish higher education institutions—is acknowledged for providing computational resources for analyses.

Data availability

Code reproducing the analyses and figures in the manuscript are available on GitHub at https://github.com/thirtysix/TFBS_footprinting_manuscript. The results of the MH vs. Neanderthal promoter analysis are available at https://osf.io/r2mtw/.

Funding

This work was supported by the Finnish Cultural Foundation and Fimlab to HB and the Academy of Finland and the Jane & Aatos Erkko Foundation to SP.

Conflicts of interest

The authors declare that they have no conflicts of interest.

References

1. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. RESEARCH ARTICLE A Draft Sequence of the Neandertal Genome. Science. 2010;328: 710–723. doi:10.1126/science.1188021

2. Castellano S, Parra G, Sanchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, et al. Patterns of coding variation in the complete exomes of three Neandertals. Proceedings of the National Academy of Sciences. 2014;111: 6666–6671. doi:10.1073/pnas.1405138111

3. Prüfer K, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358: 655–658. Available: http://science.sciencemag.org/

4. Hajdinjak M, Fu Q, Hubner A, Petr M, Mafessoni F, Grote S, et al. Reconstructing the genetic history of late Neanderthals. Nature. 2018;555: 652–656. doi:10.1038/nature26151

5. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature. 2010;468: 1053–1060. doi:10.1038/nature09710

6. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338: 222–226. doi:10.1126/science.1224344

7. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505: 43–49. doi:10.1038/nature12886

8. Petr M, Paabo S, Kelso J, Vernot B. Limits of long-term selection against Neandertal introgression. Proc Natl Acad Sci U S A. 2019;116: 1639–1644. doi:10.1073/pnas.1814338116

9. Barker H, Aaltonen M, Pan P, Vähätupa M, Kaipiainen P, May U, et al. Role of carbonic anhydrases in skin wound healing. Exp Mol Med. 2017. doi:10.1038/emm.2017.60

10. Karjalainen SL, Haapasalo HK, Aspatwar A, Barker H, Parkkila S, Haapasalo JA. Carbonic anhydrase related protein expression in astrocytomas and oligodendroglial tumors. BMC Cancer. 2018;18. doi:10.1186/s12885-018-4493-4

11. Barker H, Parkkila S. Bioinformatic characterization of angiotensin-converting enzyme 2, the entry receptor for SARS-CoV-2. PLoS One. 2020;15: e0240647. doi:10.1371/journal.pone.0240647

12. Arppo A, Barker H, Parkkila S. Bioinformatic characterization of ENPEP, the gene encoding a potential cofactor for SARS-CoV-2 infection. PLoS One. 2024;19: e0307731. doi:10.1371/journal.pone.0307731

13. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

14. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science Engineering. 2007;9: 90–95. doi:10.1109/MCSE.2007.55

15. Waskom M, Gelbart M, Botvinnik O, Ostblom J, Hobson P, Lukauskas S, et al. mwaskom/seaborn: v0.11.1 (December 2020). 2020. doi:10.5281/zenodo.4379347

16. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;25: 288–289. doi:10.1093/bioinformatics/btn615

17. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47: D419–D426. doi:10.1093/nar/gky1038

18. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019;573: 61–68. doi:10.1038/s41586-019-1506-7

19. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19: 15. doi:10.1186/s13059-017-1382-0

20. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15: e8746. doi:10.15252/msb.20188746

21. Chippada B. forceatlas2: Fastest Gephi's ForceAtlas2 graph layout algorithm implemented for Python and NetworkX. Github; Available: https://github.com/bhargavchippada/forceatlas2

22. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One. 2014;9: e98679. doi:10.1371/journal.pone.0098679

23. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14: 128. doi:10.1186/1471-2105-14-128

24. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44: W90-7. doi:10.1093/nar/gkw377

25. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45: D833–D839. doi:10.1093/nar/gkw943

26. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016;13: 966–967. doi:10.1038/nmeth.4077

27. Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol. 2021;17: e9923. doi:10.15252/msb.20209923

28. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell. 2018;175: 1701-1715.e16. doi:10.1016/j.cell.2018.10.022

29. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347: 1260419. doi:10.1126/science.1260419

30. Ding SL, Royall JJ, Sunkin SM, Ng L, Facer BA, Lesnar P, et al. Comprehensive cellular-resolution atlas of the adult human brain. J Comp Neurol. 2016;524: 3127–3481. doi:10.1002/cne.24080

31. Doan RN, Bae BI, Cubelos B, Chang C, Hossain AA, Al-Saad S, et al. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell. 2016;167: 341-354 e12. doi:10.1016/j.cell.2016.08.071

32. de Mendoza A, Sebe-Pedros A. Origin and evolution of eukaryotic transcription factors. Curr Opin Genet Dev. 2019;58–59: 25–32. doi:10.1016/j.gde.2019.07.010

33. Koonin EV. Viruses and mobile elements as drivers of evolutionary transitions. Philos Trans R Soc Lond B Biol Sci. 2016;371. doi:10.1098/rstb.2015.0442

34. Catarino B, Hetherington AJ, Emms DM, Kelly S, Dolan L. The Stepwise Increase in the Number of Transcription Factor Families in the Precambrian Predated the Diversification of Plants On Land. Mol Biol Evol. 2016;33: 2815–2819. doi:10.1093/molbev/msw155

35. Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM. Genesis and expansion of metazoan transcription factor gene classes. Mol Biol Evol. 2008;25: 980–996. doi:10.1093/molbev/msn047

36. Paps J, Holland PWH. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nat Commun. 2018;9: 1730. doi:10.1038/s41467-018-04136-5

37. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011;43: 1154–1159. doi:10.1038/ng.917

38. Erkenbrack EM, Maziarz JD, Griffith OW, Liang C, Chavan AR, Nnamani MC, et al. The mammalian decidual cell evolved from a cellular stress response. PLoS Biol. 2018;16: e2005594. doi:10.1371/journal.pbio.2005594

39. Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, et al. Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. Genome Res. 2012;22: 611–622. doi:10.1101/gr.127324.111

40. Sousa AMM, Meyer KA, Santpere G, Gulden FO, Sestan N. Evolution of the Human Nervous System Function, Structure, and Development. Cell. 2017;170: 226–247. doi:10.1016/j.cell.2017.06.036

41. Somel M, Franz H, Yan Z, Lorenc A, Guo S, Giger T, et al. Transcriptional neoteny in the human brain. Proc Natl Acad Sci U S A. 2009;106: 5743–5748. doi:10.1073/pnas.0900544106

42. Somel M, Liu X, Khaitovich P. Human brain evolution: transcripts, metabolites and their regulators. Nat Rev Neurosci. 2013;14: 112–127. doi:10.1038/nrn3372

43. Miller DJ, Duka T, Stimpson CD, Schapiro SJ, Baze WB, McArthur MJ, et al. Prolonged myelination in human neocortical evolution. Proc Natl Acad Sci U S A. 2012;109: 16480–16485. doi:10.1073/pnas.1117943109

44. Lacruz RS, Bromage TG, O'Higgins P, Arsuaga JL, Stringer C, Godinho RM, et al. Ontogeny of the maxilla in Neanderthals and their ancestors. Nat Commun. 2015;6: 8996. doi:10.1038/ncomms9996

45. Smith TM, Tafforeau P, Reid DJ, Pouech J, Lazzari V, Zermeno JP, et al. Dental evidence for ontogenetic differences between modern humans and Neanderthals. Proc Natl Acad Sci U S A. 2010;107: 20923–20928. doi:10.1073/pnas.1010906107

46. Gunz P, Neubauer S, Maureille B, Hublin JJ. Brain development after birth differs between Neanderthals and modern humans. Curr Biol. 2010;20: R921-2. doi:10.1016/j.cub.2010.10.018

47. Neubauer S, Hublin JJ, Gunz P. The evolution of modern human brain shape. Sci Adv. 2018;4: eaao5961. doi:10.1126/sciadv.aao5961

48. Banerjee-Basu S, Packer A. SFARI Gene: an evolving database for the autism research community. Dis Model Mech. 2010;3: 133–135. doi:10.1242/dmm.005439

49. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI

GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47: D1005–D1012. doi:10.1093/nar/gky1120

50. Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature. 2022;604: 502–508. doi:10.1038/s41586-022-04434-5

51. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An J-Y, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. 2020;180: 568-584.e23. doi:10.1016/j.cell.2019.12.036

52. Paciorkowski AR, Traylor RN, Rosenfeld JA, Hoover JM, Harris CJ, Winter S, et al. MEF2C Haploinsufficiency features consistent hyperkinesis, variable epilepsy, and has a role in dorsal and ventral neuronal developmental pathways. Neurogenetics. 2013;14: 99–111. doi:10.1007/s10048-013-0356-y

53. Zhou WZ, Zhang J, Li Z, Lin X, Li J, Wang S, et al. Targeted resequencing of 358 candidate genes for autism spectrum disorder in a Chinese cohort reveals diagnostic potential and genotype-phenotype correlations. Hum Mutat. 2019;40: 801–815. doi:10.1002/humu.23724

54. Nowakowska BA, Obersztyn E, Szymanska K, Bekiesinska-Figatowska M, Xia Z, Ricks CB, et al. Severe mental retardation, seizures, and hypotonia due to deletions of MEF2C. Am J Med Genet B Neuropsychiatr Genet. 2010;153B: 1042–1051. doi:10.1002/ajmg.b.31071

55. Schizophrenia Working Group of the Psychiatric Genomics, Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511: 421–427. doi:10.1038/nature13595

56. Mitchell AC, Javidfar B, Pothula V, Ibi D, Shen EY, Peter CJ, et al. MEF2C transcription factor is associated with the genetic and epigenetic risk architecture of schizophrenia and improves cognition in mice. Mol Psychiatry. 2018;23: 123–132. doi:10.1038/mp.2016.254

57. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature. 2001;413: 519–523. doi:10.1038/35097076

58. Chen YC, Kuo HY, Bornschein U, Takahashi H, Chen SY, Lu KM, et al. Foxp2 controls synaptic wiring of corticostriatal circuits and vocal communication by opposing Mef2c. Nat Neurosci. 2016;19: 1513–1522. doi:10.1038/nn.4380

59. Maricic T, Gunther V, Georgiev O, Gehre S, Curlin M, Schreiweis C, et al. A recent evolutionary change affects a regulatory element in the human FOXP2 gene. Mol Biol Evol. 2013;30: 844–852. doi:10.1093/molbev/mss271

60. Atkinson EG, Audesse AJ, Palacios JA, Bobo DM, Webb AE, Ramachandran S, et al. No Evidence for Recent Selection at FOXP2 among Diverse Human Populations. Cell. 2018;174: 1424-1435 e15. doi:10.1016/j.cell.2018.06.048

61. Sissler M, González-Serrano LE, Westhof E. Recent Advances in Mitochondrial Aminoacyl-tRNA Synthetases and Disease. Trends Mol Med. 2017;23: 693–708. doi:10.1016/j.molmed.2017.06.002

62. Fine AS, Nemeth CL, Kaufman ML, Fatemi A. Mitochondrial aminoacyl-tRNA synthetase disorders: an emerging group of developmental disorders of myelination. J Neurodev Disord. 2019;11: 29. doi:10.1186/s11689-019-9292-y

63. Guo M, Yang X-L, Schimmel P. New functions of aminoacyl-tRNA synthetases beyond translation. Nat Rev Mol Cell Biol. 2010;11: 668–674. doi:10.1038/nrm2956

64. Barcia G, Rio M, Assouline Z, Zangarelli C, Roux C-J, de Lonlay P, et al. Novel FARS2 variants in patients with early onset encephalopathy with or without epilepsy associated with long survival. Eur J Hum Genet. 2021;29: 533–538. doi:10.1038/s41431-020-00757-x

65. Duan J, Sanders AR, Moy W, Drigalenko EI, Brown EC, Freda J, et al. Transcriptome outlier analysis implicates schizophrenia susceptibility genes and enriches putatively functional rare genetic variants. Hum Mol Genet. 2015;24: 4674–4685. doi:10.1093/hmg/ddv199

66. Shi Y, Ghosh MC, Tong W-H, Rouault TA. Human ISD11 is essential for both iron-sulfur cluster assembly and maintenance of normal cellular iron homeostasis. Hum Mol Genet. 2009;18: 3014–3025. doi:10.1093/hmg/ddp239

67. Schmucker S, Martelli A, Colin F, Page A, Wattenhofer-Donzé M, Reutenauer L, et al. Mammalian frataxin: an essential function for cellular viability through an interaction with a preformed ISCU/NFS1/ISD11 iron-sulfur assembly complex. PLoS One. 2011;6: e16199. doi:10.1371/journal.pone.0016199

68. Mayr JA, Haack TB, Freisinger P, Karall D, Makowski C, Koch J, et al. Spectrum of combined respiratory chain defects. J Inherit Metab Dis. 2015;38: 629–640. doi:10.1007/s10545-015-9831-y

69. Lill R. Function and biogenesis of iron-sulphur proteins. Nature. 2009;460: 831–838. doi:10.1038/nature08301

70. Lim SC, Friemel M, Marum JE, Tucker EJ, Bruno DL, Riley LG, et al. Mutations in LYRM4, encoding iron-sulfur cluster biogenesis factor ISD11, cause deficiency of multiple respiratory chain complexes. Hum Mol Genet. 2013;22: 4460–4473. doi:10.1093/hmg/ddt295

71. Bergman O, Ben-Shachar D. Mitochondrial Oxidative Phosphorylation System (OXPHOS) Deficits in Schizophrenia: Possible Interactions with Cellular Processes. Can J Psychiatry. 2016;61: 457–469. doi:10.1177/0706743716648290

72. Jablensky A, Angelicheva D, Donohoe GJ, Cruickshank M, Azmanov DN, Morris DW, et al. Promoter polymorphisms in two overlapping 6p25 genes implicate mitochondrial proteins in cognitive deficit in schizophrenia. Mol Psychiatry. 2012;17: 1328–1339. doi:10.1038/mp.2011.129

73. Bozza M, Bernardini L, Novelli A, Brovedani P, Moretti E, Canapicchi R, et al. 6p25 interstitial deletion in two dizygotic twins with gyral pattern anomaly and speech and language disorder. Eur J Paediatr Neurol. 2013;17: 225–231. doi:10.1016/j.ejpn.2012.09.008

74. Stepanova V, Moczulska KE, Vacano GN, Kurochkin I, Ju X, Riesenberg S, et al. Reduced purine biosynthesis in humans after their divergence from Neandertals. Elife. 2021;10. doi:10.7554/eLife.58741

75. Hill GE. Mitonuclear Compensatory Coevolution. Trends Genet. 2020;36: 403–414. doi:10.1016/j.tig.2020.03.002

76. Hill GE. Mitonuclear coevolution as the genesis of speciation and the mitochondrial DNA barcode gap. Ecol Evol. 2016;6: 5831–5842. doi:10.1002/ece3.2338

77. Tobler M, Barts N, Greenway R. Mitochondria and the Origin of Species: Bridging Genetic and Ecological Perspectives on Speciation Processes. Integr Comp Biol. 2019;59: 900–911. doi:10.1093/icb/icz025

78. Soto I, Couvillion M, Hansen KG, McShane E, Moran JC, Barrientos A, et al. Balanced mitochondrial and cytosolic translatomes underlie the biogenesis of human respiratory complexes. Genome Biol. 2022;23: 170. doi:10.1186/s13059-022-02732-9

79. Couvillion MT, Soto IC, Shipkovenska G, Churchman LS. Synchronized mitochondrial and cytosolic translation programs. Nature. 2016;533: 499–503. doi:10.1038/nature18015

80. Mendez FL, Poznik GD, Castellano S, Bustamante CD. The Divergence of Neandertal and Modern Human Y Chromosomes. Am J Hum Genet. 2016;98: 728–734. doi:10.1016/j.ajhg.2016.02.023

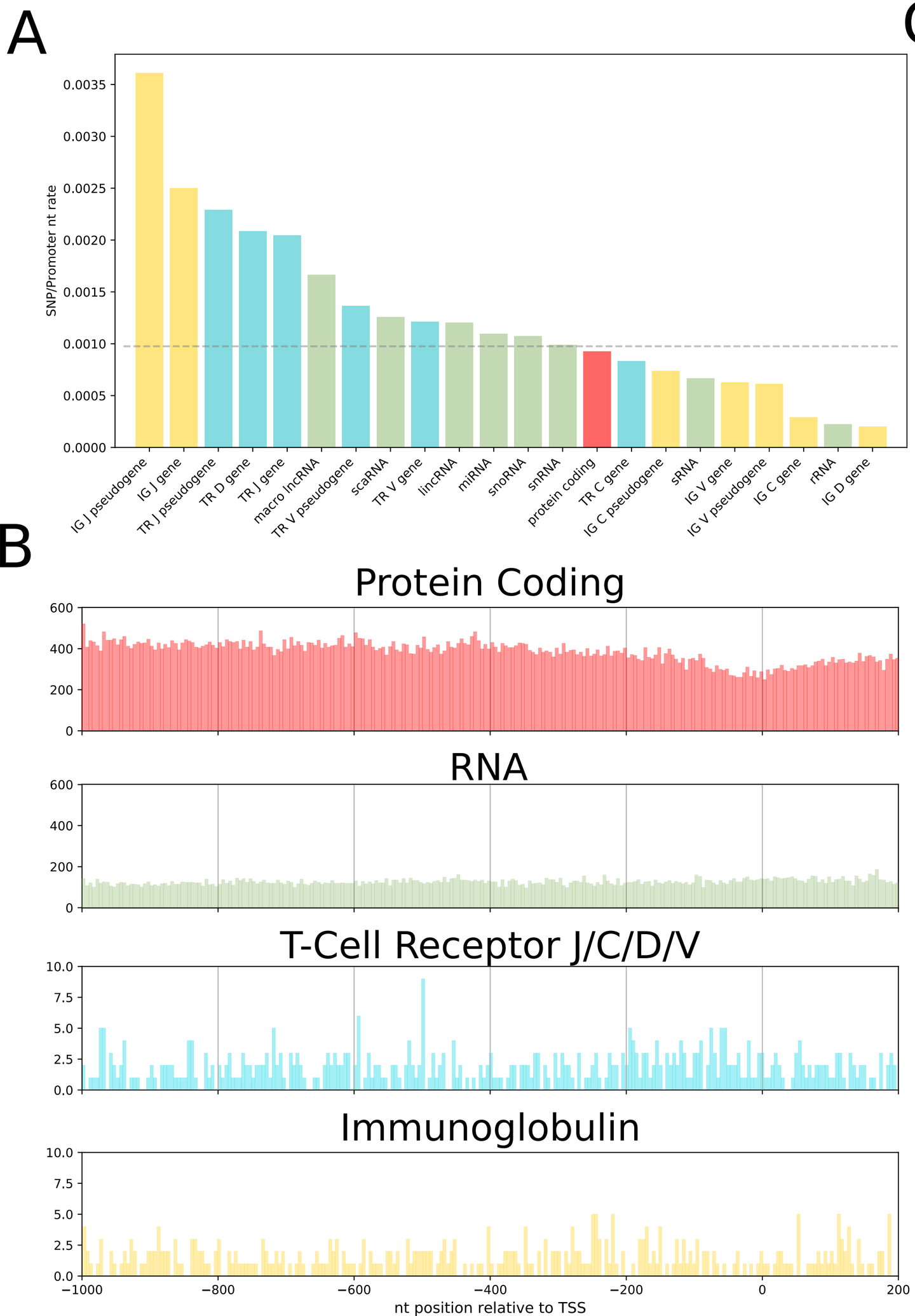
81. Endicott P, Ho SYW, Stringer C. Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of Neanderthal and modern human origins. J Hum Evol. 2010;59: 87–95. doi:10.1016/j.jhevol.2010.04.005

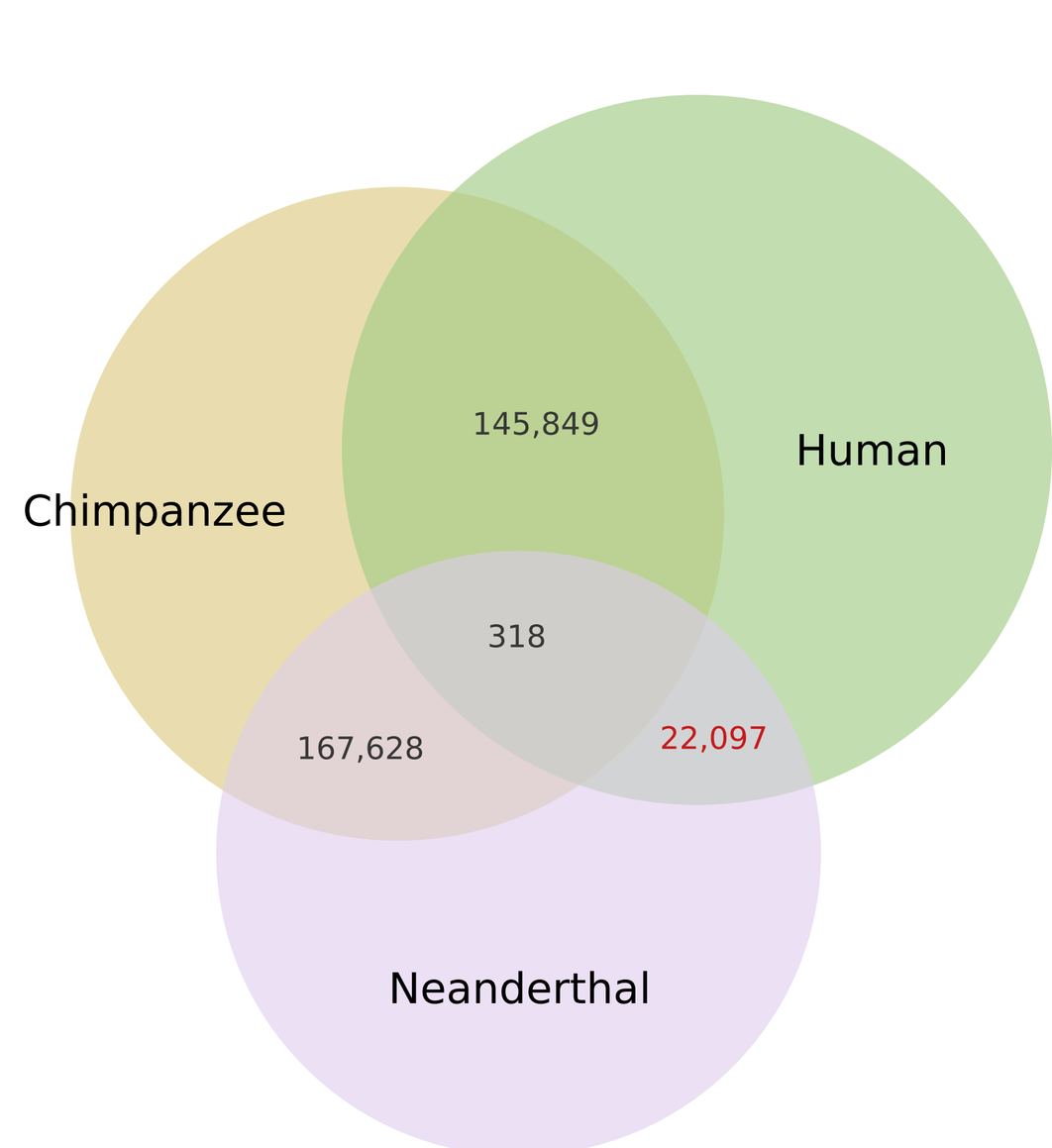
82. Gómez-Robles A. Dental evolutionary rates and its implications for the Neanderthal-modern human divergence. Sci Adv. 2019;5: eaaw1268. doi:10.1126/sciadv.aaw1268

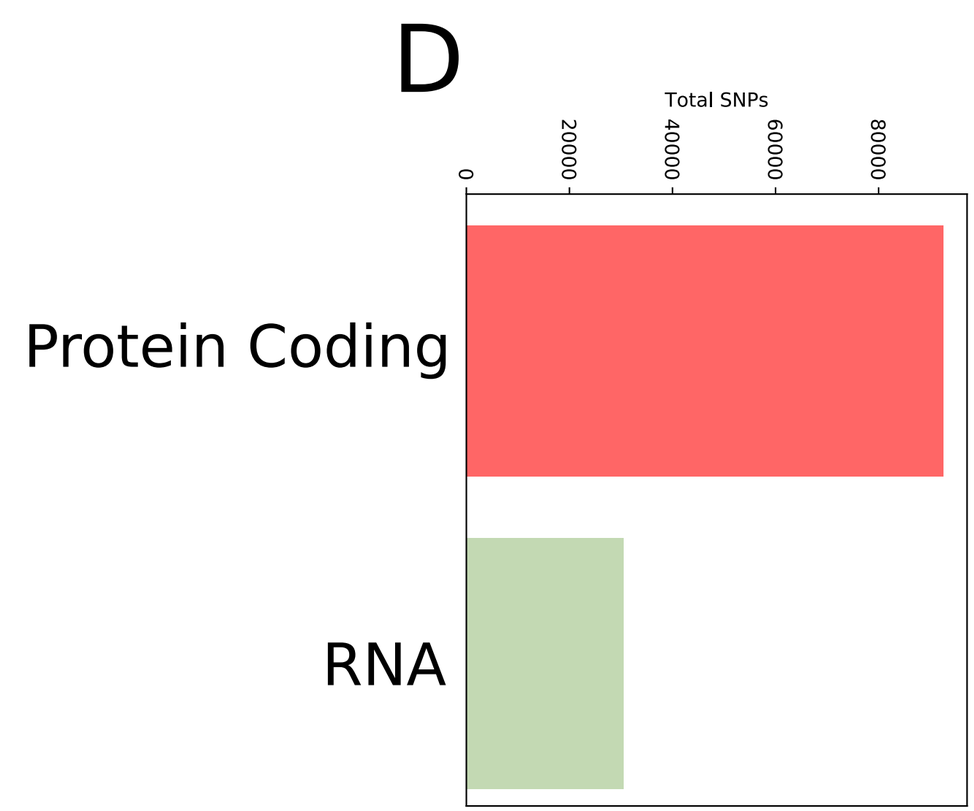
83. Zeidan J, Fombonne E, Scorah J, Ibrahim A, Durkin MS, Saxena S, et al. Global prevalence of autism: A systematic review update. Autism Res. 2022;15: 778–790. doi:10.1002/aur.2696

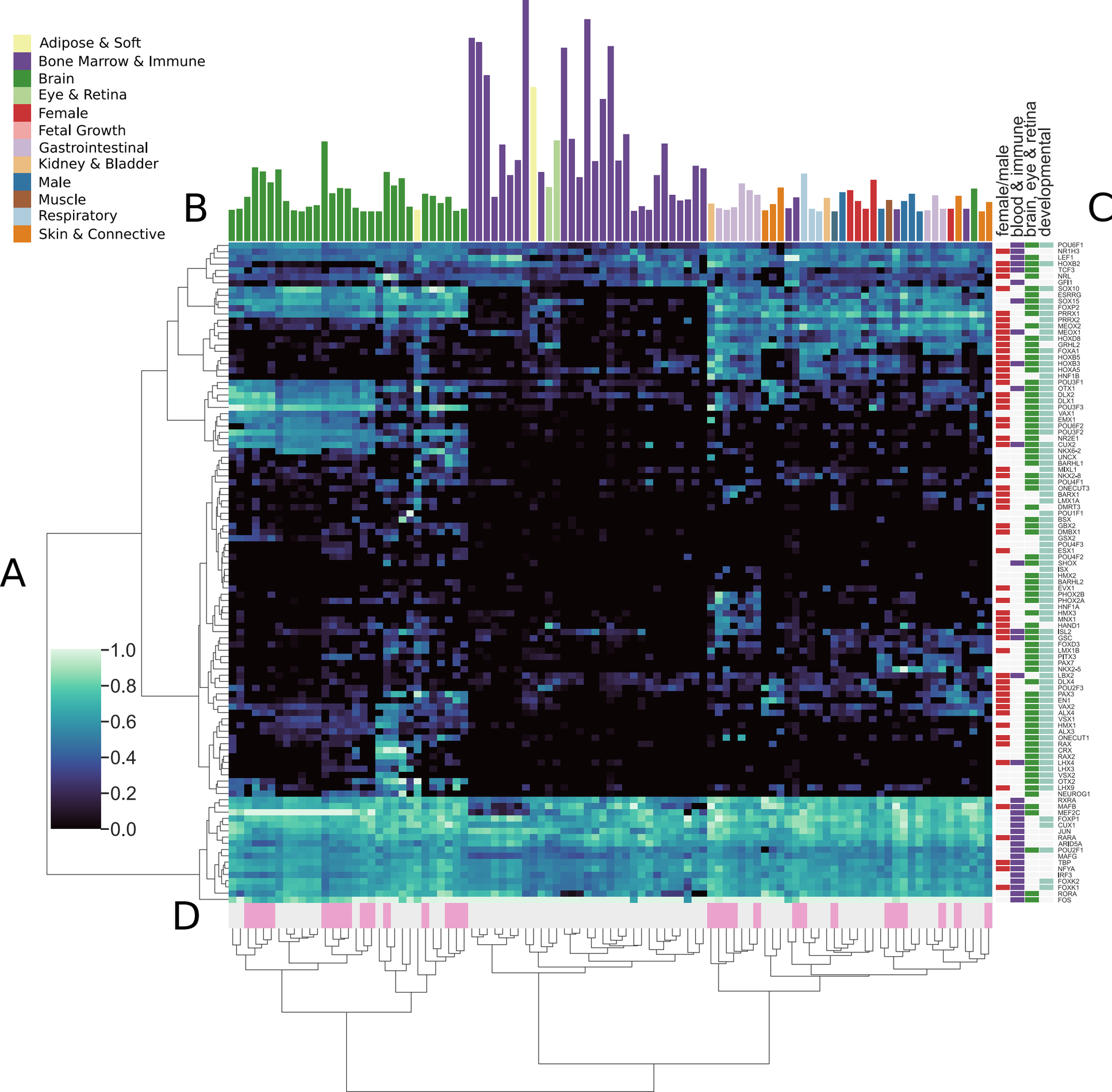
84. Charlson FJ, Ferrari AJ, Santomauro DF, Diminic S, Stockings E, Scott JG, et al. Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. Schizophr Bull. 2018;44: 1195–1203. doi:10.1093/schbul/sby058

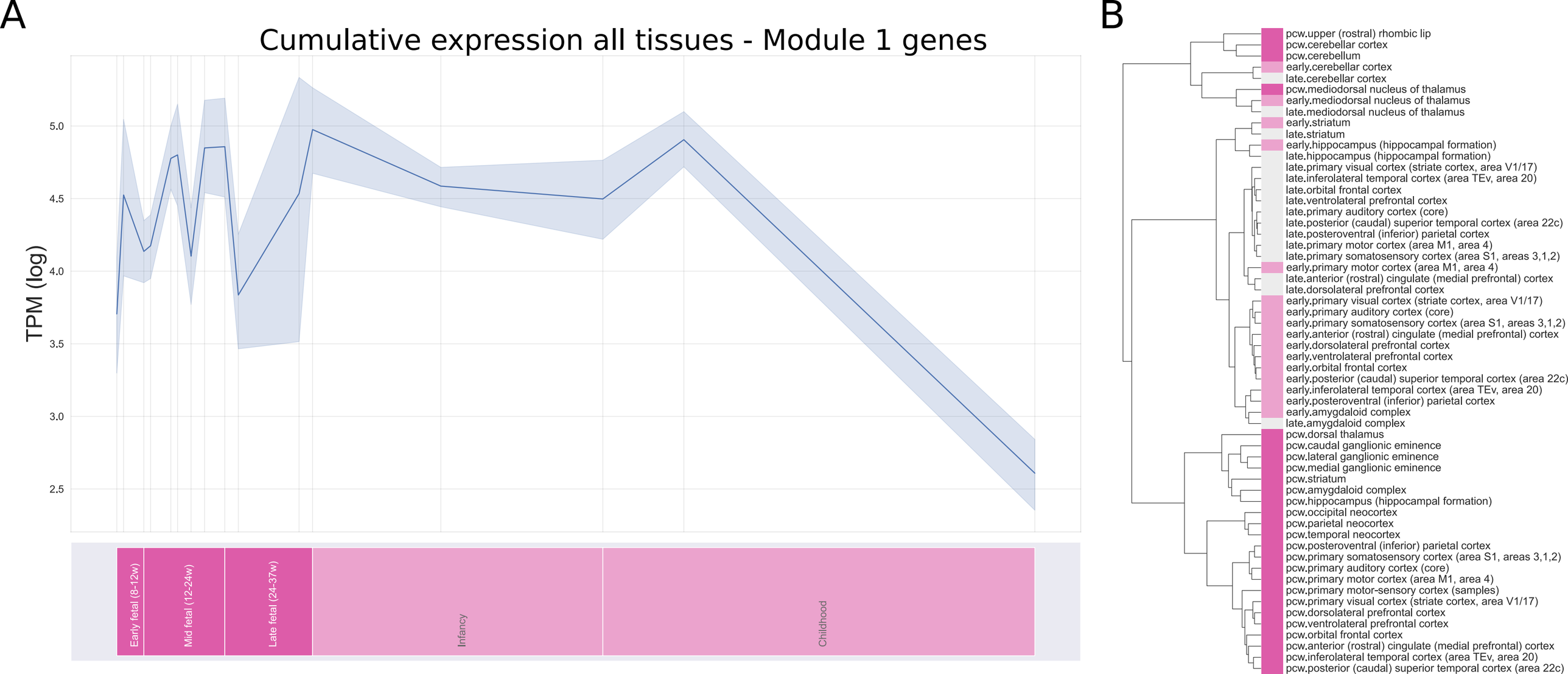
85. Linscott RJ, van Os J. An updated and conservative systematic review and meta-analysis of epidemiological evidence on psychotic experiences in children and adults: on the pathway from proneness to persistence to dimensional expression across mental disorders. Psychol Med. 2013;43: 1133–1149. doi:10.1017/S0033291712001626

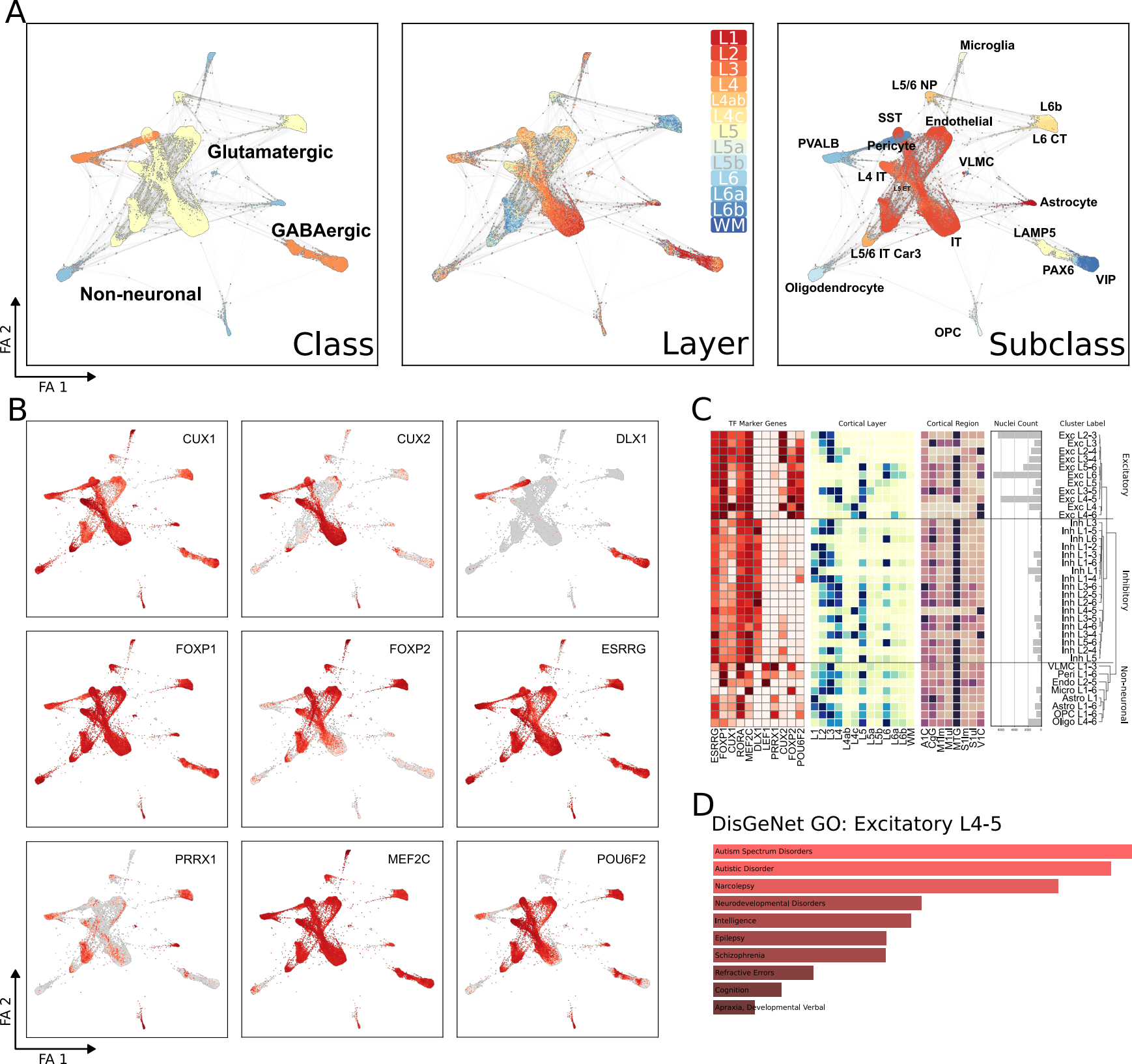

A
SNP/Promoter nt rate
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
IG J pseudogene
IG J gene
TR J pseudogene
TR D gene
TR J gene
macro lncRNA
TR V pseudogene
scaRNA
TR V gene
lincRNA
miRNA
snoRNA
snRNA
protein coding
TR C gene
IG C pseudogene
sRNA
IG V gene
IG V pseudogene
IG C gene
rRNA
IG D gene
B
Protein Coding
RNA
T-Cell Receptor J/C/D/V
Immunoglobulin
nt position relative to TSS

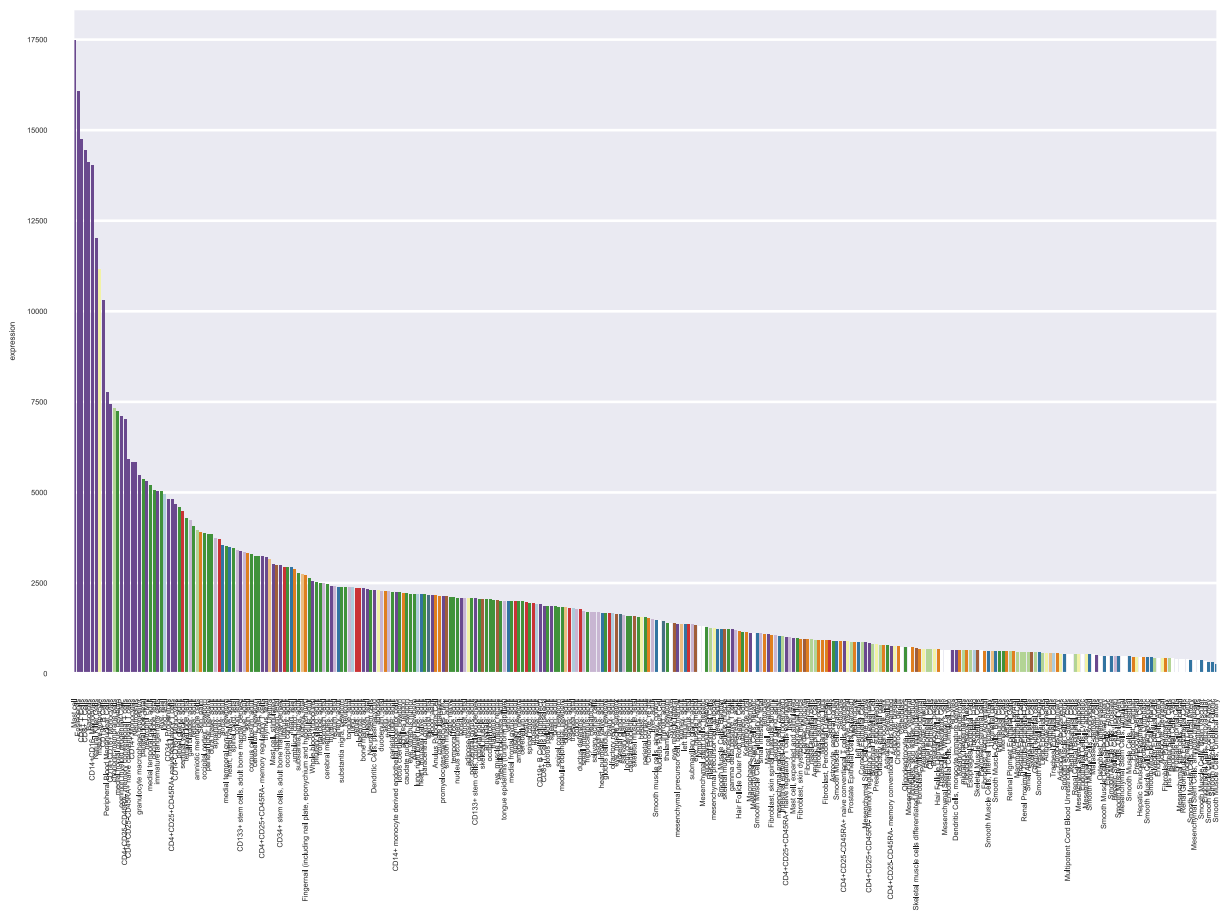

C
Chimpanzee
Human
Neanderthal
145,849
318
167,628
22,097


D
Total SNPs
0
20000
40000
60000
80000
Protein Coding
RNA


Adipose & Soft
Bone Marrow & Immune
Brain
Eye & Retina
Female
Fetal Growth
Gastrointestinal
Kidney & Bladder
Male
Muscle
Respiratory
Skin & Connective
B
C
A
D
female/male
blood & immune
brain, eye & retina
developmental
1.0
0.8
0.6
0.4
0.2
0.0
POU6F1
NR1H3
LEF1
HOXB2
TCF3
NRL
GFI1
SOX10
ESRRG
SOX15
FOXP2
PRRX1
PRRX2
MEOX2
MEOX1
HOXD8
GRHL2
FOXA1
HOXB5
HOXB3
HOXA5
HNF1B
POU3F1
OTX1
DLX2
DLX1
POU3F3
VAX1
EMX1
POU6F2
POU3F2
NR2E1
CUX2
NKX6-2
UNCX
BARHL1
MIXL1
NKX2-8
POU4F1
ONECUT3
BARX1
LMX1A
DMRT3
POU1F1
BSX
GBX2
DMBX1
GSX2
POU4F3
ESX1
POU4F2
SHOX
ISX
HMX2
BARHL2
EVX1
PHOX2B
PHOX2A
HNF1A
HMX3
MNX1
HAND1
ISL2
GSC
FOXD3
LMX1B
PITX3
PAX7
NKX2-5
LBX2
DLX4
POU2F3
PAX3
EN1
VAX2
ALX4
VSX1
HMX1
ALX3
ONECUT1
RAX
CRX
RAX2
LHX4
LHX3
VSX2
OTX2
LHX9
NEUROG1
RXRA
MAFB
MEF2C
FOXP1
CUX1
JUN
RARA
ARID5A
POU2F1
MAFG
TBP
NFYA
IRF3
FOXK2
FOXK1
RORA
FOS


A
Cumulative expression all tissues - Module 1 genes
TPM (log)
5.0
4.5
4.0
3.5
3.0
2.5
Early fetal (8-12w)
Mid fetal (12-24w)
Late fetal (24-37w)
Infancy
Childhood
B
pcw.upper (rostral) rhombic lip
pcw.cerebellar cortex
pcw.cerebellum
early.cerebellar cortex
late.cerebellar cortex
pcw.mediodorsal nucleus of thalamus
early.mediodorsal nucleus of thalamus
late.mediodorsal nucleus of thalamus
early.striatum
late.striatum
early.hippocampus (hippocampal formation)
late.hippocampus (hippocampal formation)
late.primary visual cortex (striate cortex, area V1/17)
late.inferolateral temporal cortex (area TEv, area 20)
late.orbital frontal cortex
late.ventrolateral prefrontal cortex
late.primary auditory cortex (core)
late.posterior (caudal) superior temporal cortex (area 22c)
late.posteroventral (inferior) parietal cortex
late.primary motor cortex (area M1, area 4)
late.primary somatosensory cortex (area S1, areas 3,1,2)
early.primary motor cortex (area M1, area 4)
late.anterior (rostral) cingulate (medial prefrontal) cortex
late.dorsolateral prefrontal cortex
early.primary visual cortex (striate cortex, area V1/17)
early.primary auditory cortex (core)
early.primary somatosensory cortex (area S1, areas 3,1,2)
early.anterior (rostral) cingulate (medial prefrontal) cortex
early.dorsolateral prefrontal cortex
early.ventrolateral prefrontal cortex
early.orbital frontal cortex
early.posterior (caudal) superior temporal cortex (area 22c)
early.inferolateral temporal cortex (area TEv, area 20)
early.posteroventral (inferior) parietal cortex
early.amygdaloid complex
late.amygdaloid complex
pcw.dorsal thalamus
pcw.caudal ganglionic eminence
pcw.lateral ganglionic eminence
pcw.medial ganglionic eminence
pcw.striatum
pcw.amygdaloid complex
pcw.hippocampus (hippocampal formation)
pcw.occipital neocortex
pcw.parietal neocortex
pcw.temporal neocortex
pcw.posteroventral (inferior) parietal cortex
pcw.primary somatosensory cortex (area S1, areas 3,1,2)
pcw.primary auditory cortex (core)
pcw.primary motor cortex (area M1, area 4)
pcw.primary motor-sensory cortex (samples)
pcw.primary visual cortex (striate cortex, area V1/17)
pcw.dorsolateral prefrontal cortex
pcw.ventrolateral prefrontal cortex
pcw.orbital frontal cortex
pcw.anterior (rostral) cingulate (medial prefrontal) cortex
pcw.inferolateral temporal cortex (area TEv, area 20)
pcw.posterior (caudal) superior temporal cortex (area 22c)


A
Glutamatergic
GABAergic
Non-neuronal
Class
L1
L2
L3
L4
L4ab
L4c
L5
L5a
L5b
L6
L6a
L6b
WM
Layer
Microglia
L5/6 NP
L6b
SST
Endothelial
L6 CT
PVALB
Pericyte
VLMC
L4 IT
L5 ET
Astrocyte
LAMP5
IT
L5/6 IT Car3
PAX6
VIP
Oligodendrocyte
OPC
Subclass
FA 2
FA 1
B
CUX1
CUX2
DLX1
FOXP1
FOXP2
ESRRG
PRRX1
MEF2C
POU6F2
FA 2
FA 1
C
TF Marker Genes
Cortical Layer
Cortical Region
Nuclei Count
Cluster Label
Excitatory
Inhibitory
Non-neuronal
D
DisGeNet GO: Excitatory L4-5
Autism Spectrum Disorders
Autistic Disorder
Narcolepsy
Neurodevelopmental Disorders
Intelligence
Epilepsy
Schizophrenia
Refractive Errors
Cognition
Apraxia, Developmental Verbal


expression
17500
15000
12500
10000
7500
5000
2500
0